FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Data mining na prevenção de riscos clínicos

Gonçalo Filipe Leites Pereira

Mestrado Integrado em Engenharia Informática e Computação

Orientador: João Pedro Mendes Moreira
Co-Orientador: Rui Carlos Camacho de Sousa Ferreira da Silva

8 de Outubro de 2014

Data mining na prevenção de riscos clínicos

Gonçalo Filipe Leites Pereira

Mestrado Integrado em Engenharia Informática e Computação

8 de Outubro de 2014

Resumo

Reinternamentos podem ser extremamente custosos tanto para organizações de saúde como para os pacientes. Reinternamentos podem assinalar deficiências no funcionamento de uma organização de saúde, erodem a confiança dos clientes e representam um significante acréscimo em termos custos monetários e organizacionais. Para um paciente, a necessidade reinternamento pode ser extremamente prejudicial para o seu bem estar, sendo um isto um sinal de uma falha no tratamento anterior, e podendo levar no pior dos casos danos permanentes na sua saúde. A Glintt Healthcare é uma empresa que se dedica a fornecer soluções de alta qualidade para organizações de saúde, no entanto estas não possuem componentes preditivas. Sabendo os custos de reinternamentos, foi considerado pertinente a adição desta componente através do uso de data mining. Nesta dissertação, foi desenvolvido um processo de data mining de forma a se desenvolver um modelo de classificação capaz de prever internamentos de risco que possam levar a reinternamentos. Para a criação do modelo form selecionados três técnicas distintas: Naive Bayes, Random Forests e Support Vector Machines. Foi também implementado uma técnica de classificação baseada em *Inductive Logic Programming* (ILP). Este é um sub-campo de aprendizagem de máquinas que faz uso intensivo de programação em lógica, procurnado usar conhecimento de fundo de forma a criar um conjunto de hipóteses capazes de classificar factos. Esta abordagem foi implementada como alternativa ao data mining tradicional pois consegue trabalhar com a informação sobre a estrutura relacional que se encontra nas bases de dados dos hospitais, enquanto o data mining tradicional segue um paradigma em que cada instância que queremos analisar necessita ter toda a sua informação numa linha de uma tabela, o que requer vários processos de transformação para se obter este formato que podem levar a perda de informação. Através da avaliação e comparação destas várias abordagens procurou-se o melhor modelo de de classificação para este tipo de problemas, que possa ser replicado em casos semelhantes. Tendo-se concluido que uma Random Forest de 1000 árvores apresenta os melhores resultados de todas as técnicas implementadas, com uma taxa de acerto de 68% e precisão de 70%.

Abstract

Readmissions can be extremely costly to health organizations and patients. These can signal faults in the organization's operations, erode the clients trust and represent a steep increase in monetary and organizational costs. For the patients the need for readmission can be very harmful, for this means a failure in the previous treatments, and in the worst case scenario this can cause irreparable damage to their health. Glintt Healthcare is a company which aims to provide high quality solutions for health organizations, however so far their a predictive component is absent for these solutions. Knowing the costs that readmissions have in this organizations, it was considered important the additions of a predictive components through the use of data mining.

In this thesis a data mining process was developed with the aim of creating a classification model capable of predicting risk hospitalization capable of leading to readmissions. To create the model 3 distinct data mining techniques were chosen with the aim of comparing the results: Naive Bayes, Random Forests and Support Vector Machines.

Besides the regular data mining process it was also implemented a classification technique based on Inductive Logic Programming (ILP). ILP is a sub-field of machine learning which makes intensive use of Logic Programming, making use of background knowledge to create hypothesis capable of classifying facts. This approach was implemented as an alternative to traditional data mining due to it's ability to work with data in the database relational format, while traditional mining follows a paradigm which requires each instance we want classify to be in a line of a database table, this format requires several transformation processes which may lead to the loss of information Through the evaluation and comparison of this several approaches I aimed to look for the best classification model for this kind of problems, that can be replicated in similar cases. Of all the models tested we concluded that Random Forest of 1000 delivered the best results, with 68% of accuracy and 70% precision.

Agradecimentos

Gostaria de agradecer aos meus pais o constante apoio afetivo e emocional que me deram durante o meu percurso académico, mas também à transmissão de valores culturais, os quais foram fundamentais para atingir esta importante etapa da minha vida. A todos os meus amigos, da universidade e de fora, pelos seus conselhos e bom humor que partilharam comigo. Também gostaria de agradecer aos professores João Mendes Moreira e Rui Camacho, a sua ajuda e conhecimento desta foram cruciais para a realização desta dissertação. Também os meus agradecimentos ao professor Fernando Miranda, amigo de família, que se voluntariou para tarefa ingrata de me ajudar na revisão linguística. Finalmente gostaria de agradecer à equipa do Sig da Glintt, por me terem dado esta oportunidade E pelo seu apoio incansável nos bons e maus momentos, um sincero obrigado por tudo!

Gonçalo Pereira

Conteúdo

Lista de Figuras

LISTA DE FIGURAS

Lista de Tabelas

Abreviaturas e Símbolos

AUC	Area Under Curve
ALEPH	**A** **L**earning **E**ngine for **P**roposing **H**ypotheses
BI	Business Intelligence
ILP	Inductive Logic Programming
KDD	Knowledge Discovery in Databases
LOF	Local Outlier Factor
RBF	Radial Basis Functions
RF	Random Forest
ROC	Receiving Operating Characteristics
SVM	Support Vector Machine

Capítulo 1

Introdução

Nos dias de hoje, são raras as organizações que não possuem um sistema informatizado e não façam uso de bases de dados, tal como as grandes organizações, caso de hospitais e aeroportos, que contêm grandes volumes de informação. No entanto, muitas vezes, a informação é apenas depositada nas bases de dados para consulta futura, não tirando total partido de se possuir toda a informação armazenada num repositório.

Data mining é uma disciplina que têm como objetivo a descoberta de nova informação através da análise de bases de dados já existentes, possibilitando a descoberta de padrões escondidas e prever tendências futuras. Através do uso de técnicas de data mining é possível obter informação extremamente relevante, como por exemplo detetar os padrões de compra de um cliente; prever flutuações do mercado e analisar a popularidade de produtos entre muitas outras. Consequentemente, cada vez mais organizações começam a apoiar-se nas capacidades de processamento e analise de grandes volumes de dados das técnicas de data mining como suporte para futuras decisões de negócio.

1.1 Contexto

Esta dissertação é fruto da cooperação entre a Faculdade de Engenharia da Universidade do Porto e a empresa Glintt Healthcare. A Glintt Healthcare é uma subdivisão da empresa Glintt com foco no desenvolvimento de soluções para a área da saúde. Uma das várias soluções da Glintt é o sistema de gestão de informação hospitalar, que possui uma arquitetura de Data Warehouse com diversas ferramentas para a sua exploração. Apesar deste sistema fazer uso de várias técnicas de BI (Business Intelligence) neste momento não possui uma componente preditiva, será sobre esta omissão que a esta dissertação incidirá.

Sabendo que neste momento a solução armazena uma grande quantidade de informação referentes aos aspetos administrativos e clínicos das organizações de saúde, foi considerado pertinente tentar usar técnicas de data mining de classificação de forma descobrir informação importante escondida. No âmbito desta dissertação, foi considerado prioritário a criação de um modelo preditivo para a deteção de possíveis casos de reinternamento. A criação deste modelo requererá o uso de informação significante já contida na solução, nomeadamente da informação relacionada com os dados demográficos do paciente, as intervenções médicas a que foi sujeito e dos medicamentos de consumiu. De forma a garantir a confidencialidade dos pacientes, toda informação pessoal relativa aos pacientes foi removida do conjunto de dados fornecido.

1.2 Motivação e Objetivos

Os casos de reinternamento implicam grandes custos tanto para os pacientes como para as organizações de saúde. Os reinternamentos podem significar problemas no diagnóstico e tratamento anterior, destroem a confiança da população na organização de saúde e aumentam os custos monetários e organizacionais. Devido a este facto, a implementação de uma componente preditiva de forma a detetar antecipadamente estes casos foi considerada um acréscimo de valor significante à solução já existente. As colossais quantidades de informação geradas e armazenadas diariamente pelas organizações de saúde criam um ambiente ideal para uso de técnicas de data mining. Nesta dissertação, serão implementados e otimizados vários algoritmos de pré-processamento de dados e classificação de forma a criar um modelo preditivo para os casos de reinternamento. O objetivo final desta dissertação é obter um estudo comparativo entre diversas técnicas distintas, de maneira a avaliá-las e obter o melhor candidato para implementação futura, de forma a prevenir possíveis casos de reinternamento.

1.3 Trabalho nesta área

O uso de data mining na área da saúde não é novo. As organizações de saúde já se aperceberam da utilidade de possuírem capacidades preditivas para o seu funcionamento, o que deu origem a vários estudos nesta área. Estes cobrem um grande número de temas, desde a tentativa de previsão de readmissões nas urgências, à procura de um modelo capaz de prever doenças específicas.

No entanto, as soluções de data mining dependem muito de organização para organização, pois a estrutura dos dados de uma encontra-se dependente da gestão e processos de cada organização específica. Este projeto procura comparar várias técnicas de data mining no contexto dos hospitais que usam a estrutura dos sistemas da Glintt, com foco particular no uso de ILP, uma área pouco estudada no contexto de gestão hospitalar. Isto pode ser verificado em vários trabalhos já existentes, podemos analisar o caso do artigo recente, "*Assessing the Predictability of Hospital Readmission*

Using Machine Learning Arian Hosseinzadeh"[HIV+13]. Neste artigo, os investigadores trabalharam com um conjunto de dados de grande dimensionalidade, com mais de 20 000 atributos distintos, muitos deles criados devido à necessidade de existir um atributo para cada tratamento, cada droga, e cada um dos vários elementos do sistema de saúde do Quebec. Consequentemente, devido a limitações computacionais, na etapa de preparação de dados, foram forçados a eliminar uma grande quantidade de atributos considerados pouco relevantes e a implementar métodos estatísticos de forma a criar novas atributos, como por exemplo, os mais frequentes de cada classe. O uso de ILP poderá ajudar a circunvalar estes problemas e a prevenir a perda de informação. Isto deve-se ao facto de muitas vezes este tipo de ações, na preparação de dados, serem tomados como consequência da incapacidade dos métodos tradicionais de data mining bases de dados relacionais e de necessitarem de terem toda informação de cada objeto (neste caso uma readmissão) numa única linha de uma tabela. Entre outros trabalhos já realizados nesta área encontram-se o "*Predicting Risk of Readmission for Congestive Heart Failure Patients: A Multi-Layer Approach*"[VAM+13] que procura prever os reinternamentos relacionados com falhas do coração através da implementação multicamadas dos algoritmos Naive Bayes e SVM, ao contrário desta dissertação, este artigo têm um espaço de procura menor devido aapenas tentar identificar uns reinternamentos para um problema de saúde, o que permite criar um conjunto de dados de uma tabela apenas com a informação pessoal e historial de saúde dos principais indicadores relacionados com as falhas do coração. "*A Decision Tree Model for Predicting Heart Failure Patient Readmissions*"[eSW13] é outro trabalho semelhante que procura identificar reinternamentos para falhas do coração, este usa uma abordagem baseada na criação de árvores, tal como o trabalho anteriormente mencionado, este faz uso de um espaço de procura mais especifico e têm acesso ao historial de saúde. O trabalho "*Data Mining Using Clinical Physiology At Discharge To Predict ICU Readmission*"[FCV+12] procura prever casos para unidades de cuidado intensivas através de uma abordagem de modelação difusa, usando um conjunto muito específico de indicadores de indicadores após a alta de um paciente como o batimento cardíaco e tensão arterial.

Portanto apesar de a existência de trabalhos nesta área, esta dissertação distingue-se dos referidos anteriormente através da implementação e comparação de uma grande variedade de abordagens muito distintas e ao maior espaço de procura, tentado prever todos os tipos possíveis de casos de readmissão para um hóspital. Num aspeto mais técnico é de notar que esta dissertação faz principalmente uso de dados de internamentos de carácter administrativo para prever os reinternamentos, não tendo sido disponível informação como o historial de saúde dos vários pacientes.

1.4 Ferramentas

1.4.1 Rapidminer

Rapidminer é uma ferramenta que fornece um ambiente intuitivo e robusto para tarefas de data mining, texto mining, machine learning, analítica predictiva e de negócio. Este software possui

centenas de operadores que permitem criar e acompanhar todo o processo de data mining, desde a extração de dados à validação de resultados [web]. Além disso, possui vários plugins que a permitem complementar com funcionalidades de outras ferramentas, como a Weka e a linguagem R, a principal linguagem de estatística usada nesta área. Outra vantagem do Rapidminer é a sua fácil instalação e interface intuitiva. De acordo com um inquérito realizado pelo principal website desta área, KDnuggets, Rapidminer é a ferramenta de data mining mais usada atualmente [Pia14]. Nesta dissertação foi usada o Rapidminer 5.3, a versão open-source mais recente.

1.4.2 ALEPH

ALEPH (**A** **L**earning **E**ngine for **P**roposing **H**ypotheses) é um Sistema de ILP criado como protótipo para exploração de ideias, particularmente as ideias de vínculo inverso propostas por Stephen Muggleton em 1995 [Mug95]. Desde a implementação inicial o ALEPH tem vindo a evoluir, adotando funcionalidades de vários outros sistemas ILP como CProgol, FOIL e FORS. ALEPH é implementado em Prolog e é compatível com YAP e SWi Prolog, tendo sido nesta dissertação implementado no Yap 6.2.2 Algumas das características interessantes de ALEPH em relação aos outros sistemas de ILP são: permite escolher a ordem de gerações de regras; mudar a função de avaliação e a ordem de buscas [Con08] [Sri]. Nesta dissertação foi usada a versão 5 do ALEPH.

1.5 Estrutura da dissertação

Para além da introdução, esta dissertação organiza-se em mais cinco capítulos. No capítulo 2 será introduzido o problema e descritos vários conceitos relacionados com data mining. No capítulo 3 irá ser descrita a implementação. No capítulo 4 serão descritos os resultados obtidos. No capítulo 5 conterá conclusão e trabalho futuro.

Capítulo 2

Problema e Metodologia

Neste capítulo apresenta-se o problema a tratar, a disciplina de data mining e vários conceitos, técnicas e metodologias relacionadas com data mining, como por exemplo em que consiste o processo de KDD e o *Inductive Logic Programming*.

2.1 Problema

Como já foi referido no capítulo 1, a Glintt fornece um sistema de gestão hospitalar usado por várias organizações de saúde. Este sistema apesar de fornecer dados como como indicadores de produção, de qualidade e financeiros, e informação de apoio à gestão, não possui nenhuma componente preditiva. Reconhecendo esta omissão, a Glintt propôs esta dissertação.

Para a realização deste projeto, foi fornecido uma base de dados sql server contendo um conjunto de dados reais de um hospital extraído do sistema de data warehouse da Glintt. Neste conjunto, encontra-se todo o conhecimento médico disponível relacionado com os internamentos hospitalares no período de 2012 a 2014, totalizando em 27847 internamentos distintos.

A figura 2.1 contém a estrutura do conjunto de dados fornecidos. Este está sobre a forma das seguintes tabelas de uma base de dados relacional:

- **Episódio**: Esta tabela é o identificador de cada internamento distinto, ela contém o id do episódio de internamento, as datas de entrada e saída e o id do doente;
- **Doente**: Esta tabela contém toda a informação existente na base de dados sobre o doente que foi internado, nomeadamente nacionalidade, sexo, data de nascimento,etc...;

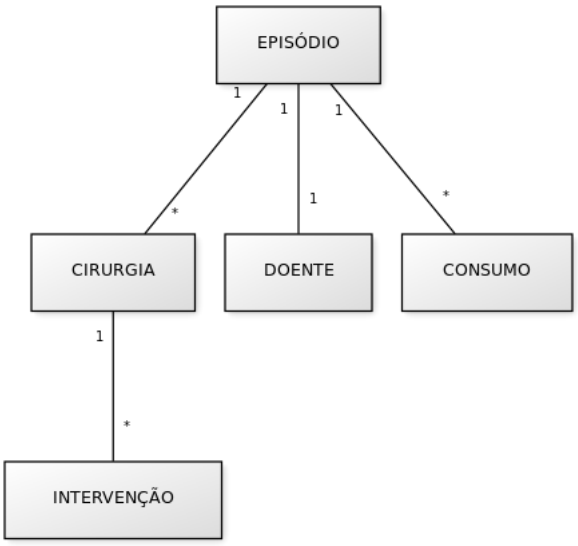

Figura 2.1: Diagrama de classes com a estrutura base de dados

- **Cirurgia**: Cada internamento pode incluir uma ou mais cirurgias. Esta tabela contém informações de quando a cirurgia foi realizada, o tempo que durou, o tipo de anestesia realizada e tempo de duração de anestesia, etc. . . ;
- **Intervenção**: Por sua vez cada cirurgia têm uma ou mais intervenção. Uma intervenção corresponde aos vários atos cirúrgicos que podem ser realizados durante uma cirurgia (e.g.remoção de apendicite, eliminação de um coágulo, etc. . .). Esta tabela contém a informação de uma intervenção, se foi a intervenção principal de uma cirurgia, e qual foi a sua posição na ordem de intervenções realizadas.;
- **Consumo**: Cada internamento pode envolver um ou mais consumos de medicamentos, cada tabela de consumo irá representar um consumo de medicamento. Esta tabela contém informações como o principio ativo do medicamento, o nome comercial, a dose tomada, o horário de consumo, etc. . . ;

De forma a explorar o conhecimento existente de forma a obtermos o modelo preditivo desejado foi seguindo o processo de *Knowledge Discovery in Databases*.

2.2 Knowledge Discovery in Databases

Com o advento da world wide web, de dispositivos de armazenamento de informação com cada vez mais capacidade e mais baratos, e a omnipresença de computadores e outros dispositivos eletrónicos como smartphones e tablets, a informação gerada por cada ser humano chegou a níveis nunca antes vistos. Para muitas organizações, conseguir analisar estas quantidades colossais de dados é extremamente crucial para os seus negócios. As tarefas usadas para descobrir novo conhecimento apartir desta informação já existente são conhecidas como tarefas de data mining. No

entanto tarefas de data mining são apenas um passo de um processo maior, conhecido por *Knowledge Discovery in Databases* (KDD) [FPSS96]. KDD representa todo o processo necessário para a extração do conhecimento, para além de data mining, isto implica outras tarefas como, a extração dos dados ou a validação dos Modelos. O processo de KDD está dividido nas seguintes partes:

- **Seleção**;
- **Pré-Processamento**;
- **Transformação**;
- **Data Mining**;
- **Interpretação/Validação**.

2.2.1 Preparação de Dados

Seleção, Pré-processamento e Transformação são considerados elementos da fase de preparação de dados. Estas etapas têm em comum o facto de envolverem todo o processo de extração e transformações do conjunto de dados antes de estes serem usados no processo data mining [FPSS96] [GCF+12].

Como um conjunto de dados pode conter ruídos, inconsistências, valores ausentes, atributos independentes sem importância para o contexto e uma quantidade de outros problemas que baixam a qualidade do conjunto. A implementação das técnicas de preparação de dados torna-se necessária para corrigir estas deficiências, especialmente em modelos preditivos mais sensíveis onde estes problemas podem fazer ou desfazer um modelo.Também é de notar que a implementação destas técnicas podem melhorar a *performance* e a adequabilidade de certas técnicas, por exemplo: a conversão de atributos polinomiais para numéricos, permite a implementação de técnicas como Support Vector Machines.

A Seleção consiste na extração dos dados que se deseja explorar. Isto não significa apenas extração direta da informação da base de dados. Esta etapa pode envolver também alguma manipulação do conjunto, como por exemplo: a eliminação manual de atributos considerados irrelevantes da base de dados, a integração de dados de vários objetos distintos num único objeto de conjunto integrado, estes muitas vezes podem até ser provenientes de fontes distintas com nomenclaturas diferentes. Nesta etapa, também, será necessário escolher a quantidade exata de objetos do conjunto de dados, apesar de teoricamente quantos mais objetos tivermos melhor serão os resultados. Isto implica, também, um grande acréscimo nos recursos computacionais necessários, é preciso alcançar um equilíbrio entre a eficiencia computacional e o desempenho do modelo resultante. Outro ponto a ter em conta, é evitar a existência de dados não-balanceados. Quando queremos classificar os dados existe a possibilidade da grande maioria dos exemplos existentes pertencerem

a uma única categoria (e.g. Numa loja que vende produtos A e B, 90% dos clientes compra A enquanto 10% compra B). Isto pode prejudicar o funcionamento dos algoritmos preditivos, criando modelos pouco fiáveis que irão ter tendência a classificar novos exemplos como pertencentes ás classes maioritárias. Devido a estas razões, nesta fase de seleção é necessário procurar balancear as classes. Para isso existem técnicas como reduzir o conjunto de objetos da classe maioritária, aumentar a quantidade de objetos da classe minoritária, atribuir custos diferentes às diferentes classes ou induzir de um modelo para uma classe.

A etapa de Pré-Processamento consiste na limpeza do conjunto de dados. Esta etapa procura melhorar a qualidade geral dos dados através do tratamento de ocorrências de ruído (que possuem valores muito diferentes do esperado), de inconsistencias (que não combinam ou contradizem outros valores do objeto), de redundancia (quando dois ou mais objetos têm os mesmos valores para todos os atributos ou dois ou mais atributos têm os mesmos valores para dois ou mais objetos) ou de dados incompletos (ausência de valores para atributos). A correção destes casos é importanto pois a presença de dados pouco fiáveis pode induzir em erro as técnicas de data mining e criar modelos de pouca confiança. Existem vários métodos possíveis para corrigir estes problemas. Por exemplo é possível em certos conjuntos de dados os utilizadores estudá-los e corrigi-los manualmente, mas obviamente esta estratégia não é viável em conjuntos de exemplos de grandes dimensões. Pode-se, portanto, definir várias estratégias para automatizar a correção destas situações. Por exemplo: nas situações em que temos casos incompletos podemos adotar três estratégias distintas: substituir os atributos em falta por pré-definidos; eliminar os exemplos com valores em falta; ou tentar criar uma regra para estimar o valor em falta. Para casos mais complexos como a deteção de ruídos, técnicas mais complexas serão necessárias, como por exemplo: a implementação de algoritmos capazes de calcular a divergência entre casos.

A etapa de Transformação consiste principalmente na alteração ou remoção de dimensões, isto deve-se ao facto, de como já foi referido, certas técnicas de data mining só conseguirem manipular certos tipos de valores (e.g. numéricos) e ao facto de o desempenho de algumas técnicas ser é influenciado pelo intervalo de variação entre os valores numéricos.

Uma das operações mais importantes nesta fase é a de Seleção de Atributos. Nesta fase, procura-se identificar os atributos mais importantes e remover os irrelevantes e redundantes com o intuito de diminuir a dimensionalidade do conjunto de dados, e consequentemente dimuniur os recursos e tempo necessarios para a sua execução, e melhorar a *performance* das várias técnicas data mining. A remoção se atributos redundantes pode melhorar a *performance*, pois alguns destes atributos podem causar *overfitting* ou induzir em erro o processo de datamining.

As técnicas desta operação estão divididas em duas categorias principais:

- **Filtros**: São técnicas aplicadas directamente conjunto de dados, procurando remover os atributos que não passam um certo critério, sem ter em consideração o impacto que irá ter na *performance* do modelo preditivo. Esta técnica é a mais eficiente computacionalmente e inclui técnicas como RELIEFF e Correlação de Pearson;

- **Wrappers**: Irão avaliar vários modelos preditivos através da remoção e/ou adicionamento de vários atributos do conjunto de dados utilizado para a sua criação, com o objetivo de encontrar a combinação ótima de dados de forma a maximizar a *performance* do modelo. Esta categoria é a que obtém melhores resultados, mas no entanto consome muitos recursos computacionais e requer muito tempo devido ao facto de necessitar de criar um modelo para cada conjunto distinto de dados. Algumas das técnicas desta categoria são o *Forward Selection* (2.3.3), *Backward Selection* e Algoritmos Genéticos [GCF+12] [MDYO04];

2.2.2 Data mining

Data mining pode ser definido como o processo de descobrir e sumarizar conhecimento útil escondido em dados históricos, de forma a ser usado como suporte na tomada de decisões. Este processo pode ser automático ou semi-automático e é multidisciplinar, envolvendo campos como aprendizagem de máquinas, reconhecimento de padrões, inteligência artificial e estatística. Os modelos resultantes deste processo inserem-se geralmente em duas categorias principais: os descritivos (não-supervisionados) e os preditivos (supervisionados) [GCF+12]. Os preditivos tem como objetivo o estudo dos dados existentes de forma a criar modelos capazes de prever um rótulo, ou valor, que caracterize um novo exemplo, com base nos valores dos seus atributos de entrada. Os descritivos têm como o objetivo descobrir padrões interpretáveis pelo ser humano, como por exemplo agrupar os exemplos em dados de conjuntos de características semelhantes. Estas duas categorias não são mutuamente exclusivas, havendo modelos que se enquadram em ambas.

Estas duas categorias ainda podem ser subdividias em mais. As tarefas descritivas podem ser divididos em 3 subcategorias: as de agrupamento, que agregam os dados de acordo com a sua semelhança; de associação, que procuram encontrar padrões de associação frequentes entre os atributos de um conjunto de dados; e de sumarização, que procura encontrar uma descrição simples e compacto para grupos de dados. As tarefas preditivas podem ser subdivididas em duas categorias: classificação e regressão. O ponto de diferenciação entre elas é o tipo de dados que elas rotulam: "classificação"irá tratar de previsões discretas, prevendo classes pré-definidas (e.g. classificação binária de Sim e Não), enquanto os de "regressão"tentam prever valores de dados contínuos (e.g. procurar prever os preços de carros).

Nesta dissertação foram usadas dois tipos de técnicas para a realização de data mining. Umas são as técnicas consideradas tradicionais, como SVM's, Naive Bayes, etc... Estas técnicas seguem um paradigma em que toda a informação para um objecto, neste caso um episódio de internamento, têm de estar contida numa linha da tabela a ser analisada. O outro tipo de técnicas usadas foram as de ILP, que serão explicadas em melhor detalhe na secção 2.3.7. Estas destacam-se no facto de conseguirem trabalhar com os dados em estrutura relacional, não sendo perdida informação através do uso processos de agregação e sumarização que muitas vezes são usadas no data mining tradicional.

2.2.3 Interpretação e Validação

Não existe uma técnica de data mining universal que permite obter o melhor desempenho para qualquer problema. Dependendo das características de cada problema, diferentes técnicas irão ter resultados distintos. Devido a esse facto, torna-se necessário a avaliação experimental.

No casos dos modelos supervisionados, são utilizadas métricas de erro baseadas na análise do desempenho do modelo criado, quando este se depara com novos exemplos não utilizados no conjunto de treino. Normalmente para obter estes conjuntos serão utilizadas técnicas de amostragem, estas irão dividir o conjunto original em vários subconjuntos de acordo com as características da técnica, havendo sempre um conjunto de treino para a criação do modelo preditivo e um conjunto de validação para a obtenção de estimativa do desempenho do preditor. De forma a avaliar os modelos de classificação estas são as técnicas principais [GCF+12]:

- Taxa de Erro: Taxa equivalente à proporção de exemplos de um conjunto de dados classificados incorretamente.
$$err(f) = \frac{1}{n}\sum_{i=1}^{n} I(y_i) \neq f(x_i)$$
Em que num conjunto de n instâncias compara-se a classe conhecida x_i, y_i, com a classe predita f.

- Matriz de Confusão: Matriz que ilustra o número de predições corretas e incorretas de cada classe. Através desta matriz é possível obter uma série de métricas de desempenho, como a precisão e a sensibilidade.

- ROC (*Receiving Operation Characteristics*): Em problemas binários é possível avaliar o problema através de um gráfico bidimensional. Estes gráficos, conhecidos como gráficos ROC, têm como eixo X a taxa de falsos positivos e como Y a taxa de verdadeiros positivos. O desempenho de um dado classificador pode ser representado por um ponto nesse espaço.

Matriz de Confusão e ROC irão ser apresentados com mais detalhe nas secções 2.3.9 e 2.3.10.

Para a avaliação dos modelos de regressão é usado tipicamente Erro Médio Quadrático.

$$EMQ(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

Em que o erro da hipótese f é calculado através da distância da classe conhecida y_i e o valor predito pelo modelo $f(x_i)$.

Neste caso, quanto menor for o valor de erro obtido melhor é a capacidade preditiva do modelo.

Após obtidos os modelos, é necessário realizar um estudo comparativo. É possível fazê-lo através da análise manual dos valores de *performance* obtidos, mas em certos casos mais difusos isso não é possível. Nessas situações, podemos usar testes como o de *Wilcoxon signed-ranks* para dois

modelos ou o teste de Friedman para mais que dois, no geral não há consenso sobre os teste mais corretos para serem implementados, portanto não serão abordados com mais detalhe.

2.3 Técnicas Utilizadas

Durante a realização de KDD irão ser usados vários algoritmos. Estes estão agrupados em duas categorias, os de preparação de dados que incluem o *Local Outlier Factor*, RELIEFF e *Forward Selection*, os de data mining que irão incluir o *Naive Bayes*, *Random Forest*, SVM e ILP. Os de preparação de dados irão modificar o conjunto de dados de forma a adotarem uma estrutura mais adequada para o processo de datamining, enquanto os classificadores irão realizar o processo de data mining em si e criar um modelo de classificação. Nesta secção é explicado o modo como estes algoritmos funcionam.

2.3.1 Local Outlier Factor

Para a deteção de *outliers* (2.2.1), foi utilizado o algoritmo *Local Outlier Factor* (LOF) [BKNS00], este faz uso do da densidade local para detetar elementos *outliers*. Este algoritmo foi desenvolvido por Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng e Jörg Sand em 2000 para a deteção de anomalias no conjunto de dados. Neste algoritmo, para verificar se objeto é um *outlier* temos que realizar as seguintes etapas:

1. Calcular a distância entre o objeto e o k^{o} vizinho mais próximo;

2. Calcular a *Reachability-Distance* do objeto. Este cálculo é feito de forma a se obter resultados mais estáveis

$$reachdist_k(p,o) = max\{kdistance(o), d(p,o)\}$$

3. Calcula-se $|N_{MinPts}(p)||$. Este valor será a quantidade de objetos de um conjunto que consiste em todos os objetos que se encontram dentro de um círculo que tem como raio a distância entre o objeto que estamos a analisar e o k vizinho mais próximo;

4. De seguida, calcula-se a *local reachability density*

$$lrd_{MinPts}(p) = \frac{1}{\left\{ \frac{\sum_{o \in N_{MinPts}(p)} reachdist_{MinPts}(p,o)}{||N_{MinPts}(p)||} \right\}}$$

5. Finalmente, calcula-se o rácio de densidades locais para o objeto

$$LOF_{MinPts}(p) = \left\{ \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{||N_{MinPts}(p)||} \right\}$$

Caso os valores obtidos sejam significantemente superiores a 1, isso significa que o valor é um *outlier*.

2.3.2 RELIEFF

RELIEFF [RSK03] é um algoritmo simples e eficiente para estimar a qualidade dos atributos e filtrá-los (2.2.1). Este irá selecionar uma instância aleatória do conjunto de dados X e procurar os K vizinhos mais próximos (H) da mesma classe e os K mais próximos da cada classe diferente (M). Para avaliar a qualidade dos atributos, será calculada a contribuição da media dos H e M usando a seguinte fórmula, onde $W[A]$, representa o peso de cada atributo,$W[A]_{-1}$ representa o peso na iteração passada em que o processo irá ser repetido N vezes com um (X) distinto

$$W[A] = W[A]_{-1} - Avg(diff(A,X,H)) + AVG(diff(A,X,M))$$

Através desta fórmula pode-se verificar que quando os H da mesma classe tiverem valores diferentes nos atributos a qualidade irá diminuir, e no caso em que nos M tiverem valores diferentes na mesma classe a qualidade irá aumentar.

2.3.3 Forward Selection

Forward Selection [KJ13] [GCF+12] é um algoritmo *wrapper* (2.2.1) usado na seleção de atributos. O Rapidminer usa uma variante deste no algoritmo no processo de seleção. *Forward Selection* irá criar uma população com n indivíduos, sendo n o número de atributos distintos no conjunto de dados. Cada um dos n indivíduos irá ser iniciado com apenas um dos atributos. Os n conjuntos serão então avaliados e aquele que apresentar o melhor resultado irá ser copiado várias vezes de forma a que seja possivel adicionar um dos atributos não selecionados a cada um dos elementos do novo conjunto de atributos. Os indivíduos do novo conjunto irão ser por sua vez avaliados e este processo irá ser repetido um sem número de vezes até não for possível melhorias nos conjuntos de atributos.

De forma a garantir algumas distinção entre a validação cruzada realizada dentro do *wrapper* e a que irá avaliar o modelo de classeficação, usou-se uma semente aleatoria distinta entre ambas, de forma a garantirmos que as partições geradas têm elementos diferentes.

2.3.4 Naives Bayes

Naive Bayes [KJ13] [GCF+12] é um classificador popular baseados no Teorema de Bayes. Este classificador usa uma assunção ingênua em que cada atributo contribui igualmente e independentemente para o resultado final e que não existem atributos latentes ou escondidos. Apesar desta simplicidade, este algoritmo geralmente consegue obter muitos bons resultados e é computacionalmente eficiente, o que o torna atrativo em vários domínios, especialmente naqueles com grande dimensionalidade.

Através do Teorema de Bayes, sabendo a probabilidade do atributo $P(x)$, da classe $P(c)$ e a probabilidade condicional $P(x|c)$ conseguimos calcular a probabilidade apriori $P(c|x))$

$$P(c|x) = \frac{P(x|c)P(c)}{P(\mathbf{x})}$$

A fórmula do classificador naive bayes é derivada desta.

$$P(c|x) \propto P(c) \prod_{n=1}^{d} P(x_n|c)$$

Sabendo que temos de classificar um objeto numa classe x, podemos considerar que a classe onde o decomposto do produto entre as probabilidades condicionais dessa classe é maior, é a classe a que este pertence. Todas as probabilidades necessárias para este classificador são obtidas a partir do conjunto de dados de treino, podendo usar técnicas como médias, variâncias ou discritização para as obter.

2.3.5 Random Forest

Random Forest [Bre99] é um classificador robusto desenvolvido por Leo Breiman e Adele Cutler. Este tem como o objetivo a criação de várias árvores de decisão de forma que, quando um novo objeto tem que ser classificado, este irá passar por as várias árvores, sendo a classificação mais comum considerada como final. Neste classificador será definido um tamanho Y para a floresta, sendo Y o número de árvores. Para cada árvore da floresta irá ser selecionado um grupo de N amostras aleatórias do conjunto de dados original que irá servir de conjunto de treino. De seguida serão também selecionados X atributos do conjunto e o melhor entre eles irá ser escolhida como split do nó. Nesta dissertação foi usada a implementação *random-input* em que este conjunto de X atributos é escolhido aleatoriamente. Esta aleatoridade é introduzida no algoritmo de forma a minimizar a correlação enquanto mantendo uma baixa taxa de erro. É importante evitar demasiada correlação entre as árvores pois, isso implica que um ou mais atributos são preditores muito fortes e que consequentemente estão a ser constantemente usados na criação de árvores semelhantes. O uso de de seleção de atributos aleatória permite que a *Random Forest* tenha boa *accuracy* e

que se mantenha robusta e resistente a *outliers* e barulho, tornado-a superior a outros algoritmos semelhantes como *Adaboost*. Este processo de ramificação irá ser repetido até a árvore gerada não poder crescer mais.

2.3.6 SVM

Support Vector Machines [KJ13] é um método de data mining supervisionado que visa em criar um hiperplano que irá dividir os exemplos em duas classes com o máximo de intervalo possível entre o hiperplano e os exemplos.

De forma a obter o hiperplano ideal, o SVM faz uso dos vetores de suporte. Os suporte suporte são os elementos nas fronteiras do conjunto de treino mais próximos do hiperplano que irão ser usados para definir a localização do hiperplano. Estes serão os elementos em que:

$$y_i(w.x_i - b) \geq 1$$

Sendo w o vetor normal ao hiperplano, x_i o objeto e b um número real. De forma a construir um hiperplano foi desenvolvido o seguinte algoritmo capaz de classificar problema não-lineares:

1. Sejam os multiplicadores de lagrange $\alpha* = (\alpha_1*, ..., \alpha_n)$ a solução de:
2. Maximizar: $\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} y_i y_j \alpha_i \alpha_j K(x_i, x_j)$
3. Sob as restrições:
$$\begin{cases} \sum_{i=1}^{n} y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, ..., n \end{cases}$$
4. O classificador é dado por: $g(x) = sng(h(x)) = sgn(\sum_{x_i \in SV} \alpha_i^* y_i K(x_i, x_j) + b*)$
5. $b^* = \frac{1}{n_{SV:\alpha^* < C}} \sum_{x_i \in SV: \alpha_j^* < C} (\frac{1}{y_j} - \sum_{x_i \in SV} \alpha_i^* y_i K(x_i, x_j))$

Neste algoritmo o $K(x_i, x)$ corresponde ao Kernel, este é uma função que recebe dois pontos no espaço como entradas e cálculo o produto escalar desses objetos no espaço de características. Kernel é importante pois, permite mapear os dados dados não-lineares recebidos numa nova dimensão de forma a torná-los linearmente separáveis de uma maneira viável e pouco custosa. Alguns dos principais Kernels são [wHcCjL10]:

- Linear: $K(x_i, x_j) = x_i^t x_j$
- Polinomial: $K(x_i, x_j) = (\gamma x_i^t x_j + r)^d, \gamma > 0$
- Função de Base Radial (RBF): $K(x_i, x_j) = exp(-\gamma ||x_i - x_j||^2), \gamma > 0$
- Sigmoide: $K(x_i, x_j) = tanh(\gamma x_i^t x_j + r)$

Em que γ , r, e d são parâmetros do kernel.

Entre as vantagens do SVM's encontram-se:

- Boa capacidade de generalização.
- Robustas perante objetos de grande dimensão.
- É definido por um problema de otimização convexo o que implica que não existem mínimos locais.
- O uso de kernel torna o algoritmo muito eficiente.

Entre as desvantagens do SVM's encontram-se [SB04] [dis]:

- Dificuldade na parametrização.
- Computacionalmente exigente o que pode prejudicar a *performance*.

2.3.7 Inductive Logic Programming

Inductive Logic Programming (ILP), é um sub-campo de data mining que faz uso intensivo de programação lógica [LD94]. Este funciona de uma maneira muito distinta dos outros métodos de data mining, particularmente devido ao facto de puderem trabalhar com a informação numa estrutura relacional, ao contrario dos outros métodos mais tradicionais que seguem o paradigma em que toda a informação precisar de estar contida num exemplo de uma tabela. ILP usa o conhecimento de fundo e exemplos de treino, de forma a criar várias hipóteses sobre a forma como os diferentes elementos do domínio se relacionam, estando tanto o conhecimento passado como os exemplos sobre a forma de factos de uma base de dados lógica. Este conjunto de hupóteses é conhecido como a teoria. A aprendizagem de ILP tem sido aplicada com sucesso em diversos tipos de problemas, como por exemplo o diagnóstico de doenças.

ILP diverge do data mining tradicional devido ao uso de uma linguagem de representação e à capacidade de usar o conhecimento do domínio. Este conhecimento é muito importante devido ao facto de permitir encontrar relações desconhecidas nos exemplos em termos das relações conhecidas do domínio. Conhecimento do domínio relevante pode permitir obtermos resultados mais precisos e eficazes no potencial conhecimento induzido, enquanto conhecimento irrelevante pode ter resultados opostos. A arte de ILP é saber selecionar e formular o conhecimento de fundo para a tarefa de aprendizagem. Devido às diferenças entre as técnicas de data mining tradicional e para o ILP, o processo de KDD será diferente para estes, sendo as diferenças elaboradas em detalhe na secção de ILP do capítulo 3

Os principais ponto fortes de ILP são [Con08]:

- Modelos complexos podem ser construídos;
- ILP usa uma linguagem de representação poderosa;

- É fácil de entender os resultados;
- É possível adicionar informação sobre o domínio;
- Há muitas domínios onde ILP pode ser usado;
- Quase todos os sistemas de ILP podem ser encontrados online;

Entre os pontos fracos incluem-se:

- Pode demorar muito tempo a alcançar resultados;
- É necessário ser um utilizador experiente de ILP para usar estes sistemas;
- O espaço de procura cresce muito rapidamente com o numero de relações do conhecimento de fundo;

O conhecimento de ILP encontra-se normalmente especificado formalmente sobre a forma de cláusulas de Horn. Uma Cláusula de Horn é uma cláusula (disjunção de n>=0 literais) que têm no máximo um literal positivo. Por exemplo, na linguagem de programação em lógica Prolog, *carta (paus,4).* é uma Cláusula de Horn em que carta será o predicado e "paus" e 4 serão os átomos que irão servir de argumentos da cláusula. Cláusulas como *carta(paus,4).* são conhecidas especificamente como factos pois, são cláusulas definidas com um literal positivo e nenhum negativo. Outros tipos de Cláusulas de Horn são as Cláusulas de Meta/Consulta que não tem nenhum literal positivo e Cláusulas de Regra que contêm um literal positivo e um ou mais literais negativos [Mov].

Nome	**Lógica Proposicional**	**Prolog**
Cláusula Facto	u	p1:-.
Cláusula Meta/Consulta	$false \leftarrow p \wedge q \wedge ... \wedge t$	:- r1, r2, ..., rm.
Cláusula de Regra	$u \leftarrow p \wedge q \wedge ... \wedge t$	p1:-q1, q2, ..., qn.

Tabela 2.1: Tipos de cláusulas de Horn

O processo de aprendizagem indutiva de conceitos, isto é, a indução de hipóteses têm como objetivo que para um conjunto de factos-exemplo, subdivididos em exemplos positivos e negativos e um domínio inicial (o conhecimento de fundo), sejam criadas hipóteses que junto com o domínio inicial sejam os capazes de satisfazer a descrição dos factos, ao mesmo tempo que satisfazem um conjunto de restrições sintáticas denominadas por *Bias*. Este processo é conhecido como processo de cobertura. *Bias* é um mecanismo empregue pelos sistemas de aprendizagem para limitar a procura de uma hipótese. É importante a definição de um *bias* em ILP, porque um bias forte pode ajudar a que o espaço de procura seja mais pequeno e a aprendizagem mais eficiente, em troca da perda de um pouco de expressividade.

Considera-se que uma boa hipótese, é aquela que cobre todos os exemplos positivos e nenhum negativo, isto é, que seja completa (cobre todos os exemplos positivos) e consistente (não cobre nenhum exemplo negativo).

$$covers(H,E) = e \in E | covers(H,e) = true$$

A elaboração de boas hipóteses para problemas de aprendizagem mais complexos irão também requerer conhecimento *apriori* significante. Este permite-nos generalizar os exemplos de uma maneira mais precisa e natural. Uma hipótese é considerada completa se em relação ao conhecimento de fundo e exemplos, todos os exemplos positivos são cobertos.

$$covers(B,H,E+) = E+$$

Uma hipótese é consistente se em relação ao conhecimento de fundo e exemplos, nenhum exemplo negativo é coberto.

$$covers(B,H,E-) = 0$$

O processo de ILP usado nesta dissertação faz uso do sistema ALPEH. O ALEPH necessita de três tipos de ficheiros com dados distinto: um ficheiro f. contendo os exemplos positivos, um ficheiro .n contendo os exemplos negativos, e um ficheiros .b com todo o conhecimento de fundo, regras e diretivas que o compilador Prolog deve usar.

Os principais elementos dos ficheiros do ALEPH são [Sri][Con08] :

- Declarações de modo : Estas declarações existem no ficheiro .b e irão descrever as relações entre os objetos e tipos de dados. Elas possuem o seguinte formato:

  mode(RecallNumber,PredicateMode).

  Em que *RecallNumber* é um um valor numérico > 1 ou * que define os limites de não-determinancia de uma forma de chamada do predicado, isto é, o número de formas alternativas que cada variável pode ter, isto pode ser exemplificado pelo exemplo do progenitor. Como sabemos que cada ser humano só pode ter 2 progenitores, só existirão 2 predicados possíveis e portanto será esse o *RecallNumber*. * é o valor normalmente aplicado quando não existe ou não se conhece o número limite de formas de um predicado. *PredicateMode* será o *template* do predicado que será usado. Os argumentos do Predicado podem pertencer a três categorias, estas são definidas pelo símbolo que aparece antes do nome dos argumentos, + significa que é uma variável de input, − significa que é uma variável de output, e # que é um valor constante. Pode-se verificar isto pelo exemplo seguinte:
  :- mode(,consumo(+medicamento,#dose,-freq)).*

Nesta declaração consumo/3, "medicamento"será o input, "dose"um valor constante e "freq"uma variavel.

- Declaração de determinação: O ALEPH usa esta declaração para determinar os predicados que podem ser usados para construir uma hipótese. Tem a seguinte estrutura:

  determination(TargetName/Arity,BackgroundName/Arity).

  O primeiro argumento é nome do predicado alvo e a sua aridade, isto é, o predicado que irá aparecer na cabeça da regra induzida. O segundo argumento consiste no nome e aridade de um predicado que pode aparecer no corpo da cláusula. Tipicamente, são feitas muitas declarações para um predicado alo, as quais correspondem aos predicados relevantes para a construção de uma hipótese. Não é possível criar uma hipótese sem ter sido declarada nenhuma determinação.

- Tipos: Tipos são o conjunto de factos que têm de existir para cada argumento do predicado. Por exemplo, para o objeto do tipo consumo, os tipos poderão ser:
  consumo(vicodin, 5).
  consumo(ibuprofen, 10).
  Os ficheiros f. e n. os exemplos estarão sobre esta forma, seguindo a estrutura do conhecimento de fundo definido pelos modos.

- Parâmetros: Estes correspondem a uma grande variedade de restrições e opções, que condicionam o funcionamento do ALEPH. Estes podem influenciar vários aspetos do ALEPH como os espaços das possíveis hipóteses a serem aprendidas, a busca realizada no espaço,etc... Estes vêm segundo o seguinte formato:

  set(Parameter,Value)

  Em que *Parameter* é o identificador do parâmetro que se quer modificar e Value o valor do parâmetro que ser quer definir.

O algoritmo de ILP usado no ALEPH funciona da seguinte maneira [Sri]:

1. Seleção de um exemplo: Escolha de um exemplo para ser generalizado, caso nenhum exista o algoritmo termina, senão continua para a etapa seguinte.
2. Construção de cláusula não especifica: Esta cláusula é construída com base nas restrições da linguagem e no exemplo selecionado na seleção do passo anterior. Esta cláusula é chamada cláusula de fundo e esta etapa é conhecida como Saturação.
3. Pesquisa: Tenta criar cláusulas mais generalizadas que a cláusula de fundo, procurando o melhores conjunto de literais da cláusula. Esta etapa é conhecida como etapa de redução.

4. Remoção: As cláusulas redundantes serão eliminadas, e a melhor será adicionada à hipótese atual. Esta etapa é conhecida remoção de cobertura. Terminando esta etapa, regressa-se à primeira etapa.

2.3.8 Validação Cruzada

Validação cruzada [GCF+12] é uma técnica de re-amostragem para a validação de um modelo. Este tipo de técnicas envolve a divisão das instâncias do conjunto de dados em dois subgrupos distintos, um que será usado para a criação do modelo e um outro para a validação do modelo criado. Nas técnicas de re-amostragem, este processo será repetido varias vezes com as *performances* resultantes a serem agregadas e sumarizadas. No processo de validação cruzada, o conjunto original irá ser dividido em K sub-grupos de tamanho idêntico [WFH11], $K-1$ dessas amostras serão usadas para treinar o modelo, sendo a amostra K usada para o validar. Este processo irá ser repetidos K vezes até todos os subgrupos terem sido usadas como grupo de validação. O valor da *performance* final serão as medias da métricas das K validações. Apesar de não haver nenhuma regra definida o valor comum usado para K é 10. Isto deve-se ao facto de se obter um balanço entre eficiência computacional e o bias, sendo neste caso o bias a diferença entre o valor estimado da *performance* e o valor real.

2.3.9 Matrizes de Confusão

A matriz de confusão [GCF+12] é um método de avaliação de bases de dados que ilustra as predições corretas e incorretas em cada classe. Nesta matriz cada elemento m_ij irá representar o número de exemplos de uma classe, em quem as colunas e linhas irão representar os elementos reais das classes e os elementos que foram classificados por um modelo preditor respetivamente.

	true S	true N
pred. S	1010	455
pred. N	958	1525

Figura 2.2: Matriz de confusão

Na figura 2.2 encontra-se um exemplo de uma matriz de confusão com uma classificação binária. Existem 1968 elementos reais da classe S em que 1010 foram classificados corretamente e 958 foram erroneamente classificados como N.D os 1465 elementos classificados como S, 455 são falsos.

A partir dos valores de uma matriz de confusão podemos obter varias métricas de desempenho. Para esta dissertação foram usadas as seguintes métricas, assumindo que VP = Verdadeiros Positivos, FP = Falsos Positivos, FN = Falsos Negativos e VN= Verdadeiros negativos.

- Precisão: Proporção de exemplos classificados corretamente para uma classe.

$$Pre = VP/(VP+FP)$$

- Sensibilidade [1] : Taxa de positivos devidamente classificados de uma classe. Indica a proporção de elementos de um classe classificados corretamente.

$$Sen = VP/(VP+FN)$$

- Taxa de acerto [2]: É a proporção de classificações corretas.

$$Exa = (VP+VN)/(VP+FP+VP+VN)$$

accuracy: 64.21%

	true S	true N	class precision
pred. S	1010	455	68.94%
pred. N	958	1525	61.42%
class recall	51.32%	77.02%	

Figura 2.3: Matriz de confusão com métricas

2.3.10 ROC

ROCs [3] [Fla10] [KJ13] [GCF+12], tal como foi previamente referido, são grafos bidimensionais em que os eixos X e Y representam a taxa de falsos positivos e taxa de verdadeiros positivos, respetivamente. Neste grafos quanto mais próximo o classificador estiver do ponto $(0,1)$, melhor são as classificações. Esse ponto é considerado paraiso do ROC em que não existe nenhum falso positivos e encontram-se todos os verdadeiros positivos, o ponto $(1,0)$ é o oposto disto e é conhecido como inferno do ROC. De forma a comparar os vários classificadores, geralmente geram-se curvas no espaço ROC a partir das previsões dos classificadores e calcula-se a AUC [4] destas. O cálculo da AUC consiste, tal como o nome indica, no cálculo de toda a área abaixo de uma curva ROC. Os valores resultantes deste cálculo enquadram-se entre 0 e 1, sendo melhores quanto mais próximos estejam do 1.

O ROC possui como vantagem a possibilidade de realizar medidas de desempenho independentes das condições e custos que podem restringir outros métodos, como por exemplo: os custos associados a classificações incorretas e distribuição de classes. No entanto, também possui várias desvantagens como o de apenas funcionar em problemas com duas classes e que na avaliação da área abaixo da curva poder ocorrer perda de informação, pois estamos a generalizar a curva e a ignorar áreas de particular interesse em certos pontos dela.

[1] *Recall* em Inglês
[2] *Accuracy*
[3] Do Inglês *Receiving Operating Characteristics*
[4] Do inglês *Area Under Roc Curve*

Capítulo 3

Implementação

Para realização do objetivo desta dissertação, foi necessário conceber setups experimentais tanto para a vertente de data mining tradicional como para ILP. Isto inclui para data mining a definição do processo de KDD e quais as técnicas que estarão envolvidas, e para ILP a definição, estrutura e regras. Nesta capitulo iremos abordar todo o processo de implementação.

3.1 Seleção de dados

De forma a colocar os dados num formato adequado para a aplicação dos processos de data mining será necessário aplicar uma série de transformações. A primeira etapa deste processo será extrair informação contida na base de dados e tentar sumarizá-la toda, de forma que cada objeto da tabela que se irá criar contenha o máximo de informação possível sobre cada episódio de internamento distinto.

Para se facilitar este processo, foi utilizado uma estratégia que consistiu em tentar sumarizar separadamente a informação de cada tabela, para depois no processo de criação da tabela final, juntar as várias tabelas através de *joins*.

A primeira ação realizada em cada tabela foi a de remover manualmente dimensões consideradas irrelevantes para o processo de data mining. Estas incluem dimensões como *flags* administrativas, atributos redundantes e informações pessoais como nome de pais ou número de Bilhete de Identidade.

3.1.1 Episódio e Doente

Os dados das tabelas Episódio e Doente foram simplesmente agregados numa única tabela devido ao facto de possuírem uma relação um para um. Cada Episódio tem apenas um doente e por isso não existe nenhuma necessidades de realizar uma operação de transformação mais complexa. Nesta nova tabela procurou-se sobre o conjunto de dados, quais os episódios que irão dar origem a reinternamentos. Dessa forma, considerou-se que todos internamentos em que no prazo de 5 dias após a alta o mesmo paciente volta a ser internado, irão ser episódios que dão origem a reinternamentos, adicionou-se ao conjunto de dados um atributo a classificar os episódios desta forma. Além disso, foi possível inferir usando a informação existente nesta tabela a idade do paciente quando foi internado e o tempo total de internamento.

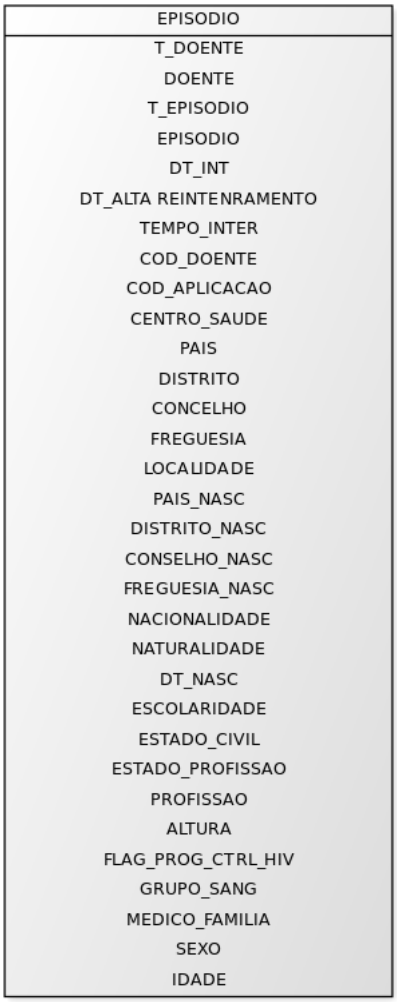

Figura 3.1: Tabela de episódios

3.1.2 Consumo

Na tabela Consumo deparamo-nos com a situação de uma relação de 1 para muitos. Como já foi referido, os métodos de data mining tradicionais requerem que toda a informação relativa a um episódio esteja contida numa única linha de uma tabela, isto pode ser problemático em situações como esta. Uma das técnicas possíveis seria pivotear a tabela de forma que cada medicamente distinto tenha uma coluna, isto é uma técnica pouco eficaz neste caso, pois não só existem centenas de medicamentos distintos, para não se perder informação iria ser necessário criar colunas extras para cada combinação de dimensões, podendo originar milhares de atributos novos, a maioria com valores em falta. No contexto desta dissertação, foram realizados testes usando essa técnica e pode-se dizer com propriedade que com a quantidade de dimensões gerada os tempos de processamento tornavam a concretização desta dissertação inviável. Por esse facto, foi adotada uma estratégia de sumarização que procura reduzir o número de dimensões novas a serem criadas. Portanto a partir das varias tabelas foi inferida a seguinte figura 3.2:

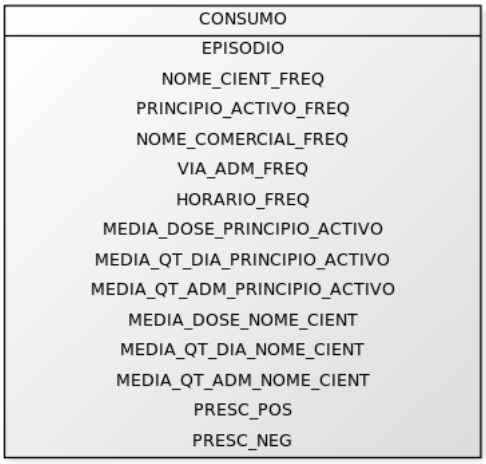

Figura 3.2: Tabela de consumos

NOME_CIENT_FREQ, *PRINCIPIO_ACTIVO_FREQ*, *NOME_COMERCIAL_FREQ*, *VIA_ADM-_FREQ*, *HORARIO_FREQ* e *ADM_FREQ* correspondem respetivamente aos valores mais frequentes do nome científico do medicamento, do princípio ativo, do nome comercial, da via de administração, do horário mais frequente e da frequência de administrações.

MEDIA_DOSE_PRINCIPIO_ACTIVO, *MEDIA_QT_DIA_PRINCIPIO_ACTIVO* e *MEDIA_QT_-ADM_PRINCIPIO_ACTIVO*, correspondem respetivamente à média da dose do príncipio ativo mais frequente, quantas vezes por dia será feita a administração dessa dose, e a quantidade de administrações totais do príncipio ativo que foram feitas. Por exemplo, para um medicamento com uma caixa de 500 mg, pode ter sido receitada uma dose de 1000 mg que será administrado em três atos diferentes ao longo do dia, o que totaliza em duas administrações totais do medicamento.

MEDIA_DOSE_NOME_CIENT, *MEDIA_QT_DIA_NOME_CIENT*, e *MEDIA_QT_ADM_NOME-_CIENT* serão dimensões idênticas às anteriores mas referente ao nome científico do medicamento em vez de princípio ativo.

PRESC_POS e *PRESC_NEG* correspondem respetivamente ao número de medicamentos consumidos que foram prescritos por um profissional de saúde e o número de medicamentes consumidos não prescritos.

3.1.3 Cirurgia

No caso das cirurgias repete-sea mesma situação de uma relação um para muitos. Nesta tabela foram escolhidos os valores mais frequentes de cada atributo para uma cirurgia de um internamento, para a média de tempo de anestesia e média de tempo cirurgia. Ao adotar esta estratégia considera-

Figura 3.3: Tabela de cirurgias

se que não há grande perda de informação em relação a internamentos com mais que uma cirurgia pois, no período estudado ocorreram 27847 internamentos, e destes apenas 1635 tiveram mais que uma cirurgia (5.9%).

3.1.4 Intervenções

Para o caso das intervenções, o problema da relação um para muitos continua presente, aqui adota-se uma estratégia ligeiramente diferente. No período estudado houve 29854 cirurgias, destas apenas 501 tiveram mais que três intervenções representando apenas 1.7% das cirurgias totais. Esta percentagem foi considerada negligenciável e daí considerarmos que uma cirurgia pode ter apenas até três intervenções distintas. Deste modo criou-se uma tabela em que cada linha corresponde ás intervenções de uma cirurgia (e consequentemente de cada episódio), com a estrutura da figura 3.4.

Em *INTERV_DESC_PRINCIPAL* e *INTERV_DT_OPER_PRINCIPAL* correspondem respectivamente à descrição e data da intervenção principal de uma cirurgia. Os conjuntos *INTERV_DESC-_SEC1*, *INTERV_DT_OPER_SEC1* e *INTERV_DESC _SEC2*, *INTERV_DT_OPER_SEC2* correspondem às descrições e datas de duas intervenções secundárias, sendo a ordem destas que decidem em que atributo são colocadas. Finalmente, também serão criados nesta tabela os atributos

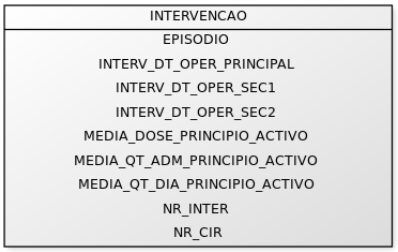

Figura 3.4: Tabela de intervenções

NR_INTER e *NR_CIR* que irão corresponder respectivamente ao número total de intervenções e cirurgias num internamento.

3.1.5 Conjunto final

Como se pode ver, todas as linhas desta tabela têm um id correspondente ao episódio de internamento, e portanto podemos juntá-las todas numa única tabela.

Tendo-se obtido a estrutura final do conjunto de dados será necessário fazer alterações a nível dos atributos.

- Nas datas de nascimento foi deixado apenas a informação do ano do nascimento, sendo considerado pouco relevantes as informções de mês e dia,
- Todos os valores relativos às datas das cirurgias foram convertidos para o dia do ano em que a cirurgia foi realizada.
- Os valores de altura irão ser discretizados em 20 pacotes distintos com gamas idênticos. Em caso de valor em falta será substiuitdo pela a gama média dos pacientes .
- Todos os valores numéricos relativos a quantidade e médias, quando em falta, foram substituídos por 0s. A todos os outros valores numéricos em falta que não pertencem a esta categoria foi-lhes atribuído o valor -1, indicando que não existem.
- Todas as strings em falta foram substituídas por uma string predefinida indicando que o valor se encontra em falta.

O último processo desta fase é o balanceamento, que foi aludido na secção 2.2.1. Em 27847 internamentos apenas 2027 irão dar origem a reinternamentos. Tanto para evitar bias na criação do modelo devido à falta de balanço como para reduzir os recursos computacionais necessários usou-se uma estratégia de subamostragem, onde se irão reduzir o número de internamentos que não originam reinternamentos. Foram selecionados 2027 membros aleatórios da classe maioritária e foi formado um novo conjunto com estes e os membros da classe minoritária. Desta forma, ficamos com um conjunto 4054 elementos em que cada classe corresponde a 50% do conjunto total de dados.

3.2 Remoção de Outliers

O próximo passo do setup é remover os casos considerados *outlier*. Nesta fase, vai-se usar o *Local Outlier Factor*(2.3.1) para procurar remover lixo do conjunto de dados. Foi implementado um operador LOF que procura utilizar 10 a 20 vizinhos mais próximos para calcular a distância Euclidiana. Sabendo que quanto mais afastado o valor resultante do LOF for de 1 mais *outlier* é, foi considerado que quaisquer valor com ≥ 2 é *outlier*. Este processo removeu 104 elementos dos valores mais afastados do conjunto de dados sem remover uma quantidade significante que possa prejudicar a criação do modelo preditivo.

3.3 Seleção de atributos

Após a remoção dos outlier passa-se para um processo de remoção de atributos. Aqui usam-se os algoritmos de filtro e um wrapper em sequência (2.2.1). No passo do filtro foi usado o RELIEFF(2.3.2). A implementação do RELIEFF usada irá ver para cada instância de um atributo os 8 vizinhos mais próximos de forma a calcular os pesos. O número de vizinhos foi obtido através de uma série de testes em que se comparou a performance do *Naive Bayes* com o conjunto de atributos originados. De acordo com o sugerido em por Mark Hall [Hal00] todos os atributos com pesos menores a 0.01 serão removidos do conjunto de dados.

Após o filtro, o conjunto de dados irá passar por o *wrapper Forward Selection* (2.3.3). Aqui, o processo de data mining diverge em três ramos distintos, um para cada algoritmo que irá ser aplicado. Dependendo do algoritmo que irá ser usado na etapa de data mining, o operador irá ser implementado de maneira diferente. Quando o algoritmo de classificação usado na etapa de data mining é uma *Random Forest* ou SVM, o algoritmo usado em *Forward Selection* irá ser uma *Random Forest* de 500 árvores. Esta decisão foi feita, visto que não era viável com os recursos computacionais disponíveis correr *Random Forests* maiores e SVMs. Como *Naives Bayes* não é um algoritmo pesado em termos de uso de recursos, foi mantido como algoritmo no *wrapper* do seu próprio ramo. Este *wrapper* irá escolher o melhor conjunto de atributos após encontrar um conjunto em que em 25 iterações do algoritmo não é possível encontrar uma combinação melhor. Após o processo de seleção de atributos o conjunto de dados irá ficar com 14 atributos.

3.4 Criação do modelo de classificação

3.4.1 Metodos Tradicionais

Estando todos os processos de preparação de dados terminados, os algoritmos de Data Mining irão ser finalmente aplicados:

- *Naive Bayes*: A implementação deste algoritmo é bastante direta, não havendo nenhum parâmetro que possa modificar os resultados finais.

- *Random Forest*: Foi implementada uma *Random Forest* com 1000 árvores. Este é um valor frequentemente usado nestas implementações, pois é suficientemente grande para a convergência das várias árvores [Bre99] num bom modelo sem ser demasiado pesado nos recursos computacionais. Foram também testados os valores menores 500, 250 e 50 para mostrar como a convergência aumenta com o número de árvores. Neste caso foi usado a implementação da ferramenta Weka, pois é mais fiel à implementação original de Breiman do que a de Rapidminer [VIM].

- SVM: Foi selecionado o C-SVM da popular implementação LIBSVM como SVM deste projeto. Foram selecionados como Kernels para esta dissertação o RBF e o Sigmoide. RBF é geralmente a primeira escolha em termos de Kernels [wHcCjL10], por ser bom a lidar com os casos em que a relação entre classes e atributos é não linear. Além disso possui menos hiperparametros que podem influenciar o resultado final e no geral menos dificuldades númericas. Como alternativa ao RBF também foi usado o sigmoide. Testes inicaís demonstraram que devido às características do conjunto de dados os kernels lineares e polinomiais têm tempos de execuções que foram considerados demasiado elevados.

  LIBSVMs têm dois parâmetro que têm-se de se procurar otimizar, C e γ. Os valores padrão destes parâmetros são $C = 1$ e $\gamma = 1/numero\_de\_features$ [CL11]. Estes valores costumam dar bons resultados, mas decidiu-se também testar uma larga gama de outros valores para estes atributos numa tentativa de se obter melhores resultados.

  É importante notar que os SVMs não aceitam valores em formato polinomial, consequentemente necessário transforma-los em valores numéricos, foi realizado com esse intuito um processo transformação qeue cria para cada valor distinto de cada atributo polinomial, um atributo novo na tabela. Cada exemplo terá no atributo novo correspondente ao valor do atributo da tabela original o valor 1 e 0 nos outros. Apesar de isto aumentar a dimensionalidade, o SVM para este conjunto de dados em particular tem tempos de execução considerados aceitáveis.

Todos estes algoritmos foram avaliados na parte final usando uma validação cruzada de 10 amostras, sendo usada a mesma semente aleatória para a geração das amostras.

3.4.2 ILP

O processo de ILP desta dissertação diverge do de data mining tradicional após a etapa de remoção *outliers*. Assim que se obtenha o conjunto de dados pós-LOF, iremos selecionar os episódios de internamento e extrair toda a informação da base dados fornecida relativa a esses episódios na estrutura relacional originalmente fornecida. Usamos o mesmo conjunto de episódios nas duas

abordagens, com o intuito de mantermos consistência entre ambas, isto é importante para o processo de avaliação. Foram também extraídos vários valores estatísticos que foram usadas na parte de data mining tradicional e que não existiam originalmente na base de dados. Toda esta informação obtida foi convertida para o formato Prolog em que cada objeto da base de dados irá ser convertido num facto, o exemplo da tabela 3.1 será convertido em:

EPISODIO	**N_REG_OPER**	**DESCRICAO_ANESTESIA**
1044305	1031830	Geral: Outras
COD_ASA	**TIPO_ANESTESIA**	**TIPO_CIR**
5	Geral: Balanceada	L
TIPO_ASSEPSIA	**QTD_DURACAO_ANESTESIA**	**QTD_DURACAO_CIRURGIA**
LIMPA	1680	1440

Tabela 3.1: Exemplo de cirurgia

cirurgia(1044305,1031830,"geral: outras",5,"geral: balanceada",l,limpa,1680,1440).
Esta informação irá servir de conhecimento passado e estará incluída no ficheiro internamento.b, os episódios positivos e negativos foram também divididos, entre os ficheiros internamento.f e internamento.b, encontrando-se no seguinte formato:
internamento(1044305).
Em que id corresponde ao id do episódio de internamento. Como se pretende validar o nosso modelo através de validação cruzada, por razões de consistência, é importante certificar-se que as partições são idênticas às de data mining tradicional. Por esse facto foram criados ficheiros para 10 conjuntos de treino e teste usando a mesma semente aleatória que em data mining tradicional.

Estando esse conhecimento criado é necessário definir as declarações de modo 2.3.7 e determinação referentes a eles no ficheiro *internamento.b*. A aridade nas declarações de determinação irá corresponder ao numero de atributos de cada tabela, enquanto os tipos do predicado de modo serão os atributos em si. Através dos inputs e outputs procurou-se simular a estrutura relacional existente.

Declarações de determinações:

```
1 :-determination(internamento/1, cirurgia/9).
2 :-determination(internamento/1, intervencao/6).
3 :-determination(internamento/1, consumo/13).
4 :-determination(internamento/1, episodio/32).
5 :-determination(internamento/1, estatistica/21)
```

Declaração de modo para internamentos:

```
1 :-modeh(1,internamento(+episodioid)).
```

Declaração de modo para episódios:

```
:-modeb(1,episodio(
       -altura
      ,-tempointer,
      ,-t_doente
      ,-t_episodio,
      ,+episodioid
      ,-dt_int
      ,-dt_alta
      ,-cod_doente
      ,-centro_saude
      ,-pais
      ,-distrito
      ,-concelho
      ,-freguesia
      ,-localidade
      ,-pais_nasc
      ,-distrito_nasc
      ,-concelho_nasc
      ,-freguesia_nasc
      ,-nacionalidade
      ,-naturalidade
      ,-escolaridade
      ,-estado_civil
      ,-estado_profissao
      ,-profissao
      ,-flag_prog_ctrl_hiv
      ,-grupo_sang
      ,-medico_familia
      ,-sexo
      ,-dt_nasc
      ,-doente
      ,-cod_aplicacao
      ,-idade)).
```

Declaração de modo para cirurgias:

```
:-modeb(*,cirurgia(
      +episodioid
      ,-n_reg_oper
      ,-descricao_anestesia
      ,-cod_asa
      ,-tipo_anestesia
```

```
      ,-tipo_cir
      ,-tipo_assepsia
      ,-qtd_duracao_anestesia
      ,-qtd_duracao_cirurgia
 )).
```

Declaração de modo para intervenções:

```
:-modeb(*,intervencao(
      +n_reg_oper
      ,-descricao
      ,-codificacao
      ,-flag_principal
      ,-n_ord
      ,-dt_oper)).
```

Como existem dezenas de milhares de consumos relativos a estes episódios, o que é extremamente elevado, e o facto de um único internamento poder ter dezenas de consultas, torna-se inviável computacionalmente termos um *RecallNumber* de *, por isso foi definido como 20. Então a declaração de modo para consumos será:

```
:-modeb(20,consumo(
      +episodioid
      ,-qt_adm
      ,-qt_dia
      ,-administracao
      ,-nome_cient
      ,-nome_comercial
      ,-principio_activo
      ,-dose
      ,-via_adm
      ,-freq
      ,-unid_med
      ,-horario
      ,-prescrito)).
```

Declaração de modo para estatísticas:

```
  :-modeb(*,estatistica(
    -interv_dt_oper_principal,
    -interv_dt_oper_sec1,
    -interv_dt_oper_sec2,
    -media_dose_principio_activo,
    -media_qt_adm_principio_activo,
```

```
7     -media_qt_dia_principio_activo,
8     -media_dose_nome_cient,
9     -media_qt_dia_nome_cient,
10    -media_qt_adm_nome_cient,
11    -media_qtd_duracao_anestesia,
12    -media_qtd_duracao_cirurgia,
13    +episodioid,
14    -nome_cient_freq,
15    -principio_activo_freq,
16    -nome_comercial_freq,
17    -via_adm_freq,
18    -horario_freq,
19    -adm_freq,
20    -interv_desc_principal,
21    -interv_desc_sec1,
22    -interv_desc_sec2)).
```

Também serão criados variantes destes modos, com uma constante (representada por # em vez -) em argumentos considerados importantes, permitindo a criação de hipóteses com valores constantes. Os argumentos considerados constantes foram: nacionalidade, sexo, país de nascimento para Episodio; tipo de cirurgia, tipo de assepsia e descrição da anestesia para Cirurgia; descrição e codificação para Intervenção; frequência, se foi prescrito e princípio activo para Consumo; a descrição da intervenção principal, o conjunto do princípio activo mais frequente com a média da sua dose, apenas o principio activo mais frequênte, conjunto do nome científico mais frequênte com a média da sua dose e apenas o nome científico mais frequente para Estatistica.

Estando os modos estabelecidos, é necessário criar várias regras auxiliares. Estas são necessárias, pois sem regras auxiliares o ALEPH poderá não conseguir generalizar o suficiente, criando hipótes demasiado específicas. Isto pôde ser verificado quando foi executado o processo de ILP sem estas regras, pelo facto de ter sido criada uma hipótese para cada exemplo na teoria.

A partir do estudo do conjunto de dados, foram inferidas as seguintes regras:

- Para os valores numéricos reais, foram defenidas regras de $\leq$ e $\geq$, de forma a encontar valores fronteira que possam dar origem a reinternamentos

```
1
2  gteq(X,Y):-
3    number(X), number(Y),
4    X >= Y, !.
5  gteq(X,X):-
6    number(X).
7
8  lteq(X,Y):-
9    number(X), number(Y),
10   X =< Y, !.
```

```
lteq(X,X):-
  number(X).
```

- Para valores que se encontram sobre a forma de texto, mas que podem representar uma escala, foi criado um conjunto de regras que representam essa situação com regras para maior ou igual e menores iguais, por exemplo, com esta regra no caso dos horários de consumos de medicamentos, o Aleph saberá que tomar "2 xs dia"(2 vezes por dia) é menor que "4 xs dia"(4 vezes por dia).

```

gteqfreq(F,E):-
consumo(_,_,_,C,_,_,F,_,_,E,_,_,_),
consumo(_,_,_,D,_,_,F,_,_,G,_,_,_),C\=D,gteqf(E,G).

lteqfreq(F,E):-
consumo(_,_,_,C,_,_,F,_,_,E,_,_,_),
consumo(_,_,_,D,_,_,F,_,_,G,_,_,_),C\=D,lteqf(E,G).

freqtier1('toma unica').
freqtier2('1 x dia').
freqtier3('2 xs dia').
freqtier4('3 xs dia').
freqtier5('4 xs dia').
freqtier6('4/4 h').
freqtier7('2/2h').
freqtier8('perfusao').


gteqf(A,B):-freqtier1(A),freqtier1(B).
gteqf(A,B):-freqtier2(A),freqtier1(B).
gteqf(A,B):-freqtier3(A),freqtier1(B).
.......
.......
lteqf(B,A):-freqtier7(A),freqtier7(B).
lteqf(B,A):-freqtier8(A),freqtier7(B).
lteqf(B,A):-freqtier8(A),freqtier8(B).
```

- Foi criado uma regra que indica se um paciente é de fora do distrito, isto pode ser importante, pois pode significar que a condição de saúde é de gravidade suficiente que não pode ser tratada localmente.

```

foradistrito(A):- episodio(_,_,_,_,A,_,_,_,_,_,_,B,_,_,_,_,_,_,_,_,_,_,_,_,
    _,_,_,_,_,_,_,_),
B \='braga',
```

```
B \=’desconhecido’,!.
```

- Um caso em que a cirurgia ao paciente tenha sido realizado no mesmo dia do internamento, pode indicar urgência no seu tratamento, foi criado uma regra que assinala este casos.

```
nodia(A):- episodio(_,_,_,_,A,B,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_
    ,_,_,_,_),cirurgia(A,C,_,_,_,_,_,_,_),intervencao(C,_,_,_,_,E),
B == E,
B \=’desconhecido’,!.
```

Para concluir a implementação de ILP foram definidos os seguintes parâmetros de execução:

- Definiu-se que um símbolo de predicado não pode aparecer mais que uma vez numa cláusula, já que apenas irá haver um atributo que servirá de chave.
- Definiu-se que para uma cláusula de hipóteses ser definida tem que cobrir pelo menos 10 exemplos positivos de forma a garantir-se que a hipótese é solida.
- Definiu-se que sempre que o processo de indução for selecionar um exemplo para generalizar, irá selecionar. três escolhidos aleatoriamente e só adicionar a melhor clausula à teoria, esta saturação de exemplos irá permitir obter-se melhores clausulas.
- Definiu-se o limite superior de nódulos quando se procura pela melhor clausula como 100000.
- Definiu-se que uma clausula de hipótese pode aceitar até 10 exemplos negativos antes de ser descartada, isto irá criar uma margem de erro aceitável para as hipóteses .

```
:- set(language, 1).
:- set(minpos, 10).
:- set(samplesize, 3).
:- set(nodes, 100000).
:- set(noise, 10).
```

3.5 Resumo

Neste capítulo foi discutida a implementação da dissertação. O conjunto de dados recebido teve de sofrer grandes alterações para ser compatível com o processo de data mining. No novo conjunto, os *outliers* foram removidos através do LOF, e os atributos desnecessários através do uso em conjunto de RELIEFF e *Forward Selection*. Estando a preparação dos

dados concluídas, o processo diverge em três caminhos para cada algoritmo que se deseja testar: *Naive Bayes*, *Random Forest* e SVM. O processo de ILP divergiu do processo de data mining após a remoção de *outliers*, este irá manter a estrutura relacional do conjunto de dados original mas usando apenas os mesmos episódios de internamento que o processo de data mining. Foram criadas várias regras auxiliares em Prolog de forma a facilitar o processo de generalização.

Capítulo 4

Resultados

Estando todos os processos implementados, podemos criar os modelos de classificação. Neste capítulo iremos comparar os vários modelos obtidos através do uso de diversas métricas de avaliação. Irá procurar-se estabelecer qual o melhor modelo criado com especial atenção na comparação dos modelos gerados tradicionalmente com o modelo gerado em ILP.

4.1 Avaliação dos processos de data mining tradicional

Esta secção estará dividia em 4 partes. As 3 primeiras partes abordaram os resultados das 3 técnicas em que é necessário parametrização: *Random Forests*, SVMs com kernel RBF e SVMs com kernel sigmoide. Na quarta parte, iremos comparar os quatro métodos de data mining tradicional e discutir os resultados. Nesta estudo foram considerados 4 métricas: Taxa de acerto, Precisão e Sensibilidade que são obtidas a partir de matrizes de confusão e AUC que é obtido a partir dos gráficos ROC.

4.1.1 Avaliação de Random Forests

Técnica	**Taxa de acerto**	**Precisão**	**Sensibilidade**	**AUC**
Random Forest - 1000 árvores	68.00%	66.94%	70.81%	0.739
Random Forest - 500 árvores	67.97%	67.00%	70.51%	0.739
Random Forest - 250 árvores	67.75%	66.89%	69.95%	0.737
Random Forest - 100 árvores	67.72%	66.76%	70.25%	0.737
Random Forest - 50 árvores	67.52%	66.72%	69.59%	0.733

Tabela 4.1: Melhores resultados de *Random Forest*

Pode-se constatar que como já foi indicado, que quanto mais árvores a *Random Forest* tiver melhor a sua capacidade de classificação. Isto deve-se ao simples facto que, quantas mais árvores são criadas, mais classificações teremos e maior a probabilidade das suas classificações convergirem num resultado correto, diminuindo a margem de erro. Também se pode constatar que, á medida que o número de árvores aumenta o melhoramento da *performance* irá diminuir, eventualmente chegando a um limiar onde os recursos usados não justificam o ganho de *performance* [BPO12]. Devido à convergência dos resultados *performance* num valor quanto mais árvores temos, o facto da precisão para 500 árvores ter sido ligeiramente superior do que para 1000 árvores, foi considerado uma variação negligível consequente da aleatoriedade na criação do modelo, já que os valores de precisão entre os dois modelos são muito próximos.

4.1.2 Avaliação dos SVMs RBF

Técnica	**Taxa de acerto**	**Precisão**	**Sensibilidade**	**AUC**
Default	61.04%	67.71%	41.83%	0.640
C=50 ;γ=0.005	62.96%	62.19%	65.63%	0.667
C=1 ;γ=0.75	57.85%	54.55%	92.74%	0.641

Tabela 4.2: Melhores resultados de SVMs com kernel RBF

Para a avaliação dos SMVs com o Kernel RBF, foram selecionados os modelos encontrados com as melhores parametrizações para cada métrica. A parametrização (C=50 ;γ=0.005) produz os melhores resultados de todas, tendo a melhor taxa de acerto e AUC de todos os testes, além disso apresenta bons resultados na precisão e sensibilidade. A parametrizão default (C=1; $\gamma = \frac{1}{numero\_de\_atributos}$) apresenta a melhor precisão de todos, no entanto a sensibilidade é muito baixa, sendo uma fraca opção para uma situação que seja prioritária, a deteção de possiveis casos de reinternamentos. Finalmente (C=1 ;γ=0.75) apresenta resultados extremamente bons em termos de sensibilidade, detectando 92.74% dos casos que geram reinternamentos. Nesta situação, o melhor resultado dependerá das prioridades do utilizador, globalmente, a parametrização (C=50 ;γ=0.005) apresenta os melhores resultados, no entanto caso a prioridade seja apenas dar o alarme relativamente possíveis casos que geram reinternamentos, ignorando o facto da possibilidade de uma grande margem de erro (C=1 ;γ=0.75) é a melhor parametrização.

4.1.3 Avaliação de SVMs sigmoide

Técnica	Taxa de acerto	Precisão	Sensibilidade	AUC
Default	53.97%	53.77%	55.08%	0.535
C=1 ;γ=0.75	52.51%	51.81%	68.43%	0.499
C=200 ;γ=10	54.30%	55.31%	43.65%	0.485

Tabela 4.3: Melhores resultados de SVMs com kernel sigmoide

Para a avaliação dos SMVs com o Kernel sigmoide, tal como para RBF, foram selecionados também os casos com os melhores valores para cada métrica. Neste Kernel, tal como se pode ver na tabela 4.3, as melhores parametrizações apresentam valores muito próximos uma das outras, havendo apenas pequenas variações entre estes. Devido a essa razão é dificil determinar qual a melhor parametrização, pois cada uma apenas apresenta ligeiras vantagens em relação a outros em cada métrica.

4.1.4 Comparação entre técnicas

Técnica	Taxa de acerto	Precisão	Sensibilidade	AUC
Naive Bayes	65.22%	57.21%	67.97%	0.701
Random Forest - 1000 árvores	68.00%	66.94%	70.81%	0.739
SVM RBF (C=50 ;γ=0.005)	62.96%	62.19%	65.63%	0.667
SVM RBF (C=1 ;γ=0.75)	57.85%	54.55%	92.74%	0.641

Tabela 4.4: Melhores resultados das várias técnicas de data mining tradicional

Nesta secção, iremos avaliar os vários métodos de data mining tradicional distintos. Primeiro, é importante notar que os resultados do SVM com kernel sigmoide não foram considerados, isto deve-se ao facto do kernel RBF superar globalmente os resultados do sigmoide. Considerou-se que isto deve-se ao facto do kernel do RBF conseguir dividir o espaço melhor do que o Sigmoide para este conjunto de dados. Isto não significa que o Sigmoide apresentará sempre piores resultados do que o RBF, mas apenas que para este conjunto de dados em particular o RBF adequa-se melhor. Pode-se verificar na tabela 4.4, que apesar dos bons resultados apresentados *Naive Bayes* e SVM RBF (C=50 ;γ=0.005), a *Random Forest* de 1000 árvores supera-os a todos, apresentado os melhores resultados para todas as métricas avaliadas. A unica excepção é o SVM RBF (C=1 ;γ=0.75) que apresenta uma melhor sensibilidade. Deste modo, pode se considerar a *Random Forest* de 1000 árvores o melhor modelo de classificação para casos de possíveis reinternamento dentro dos métodos de data mining tradicionais, apresentando excelente resultados em todas as métricas, e mesmo tendo menor sensibilidade que o SVM RBF (C=1 ;γ=0.75), 70.81% continua a ser um resultado muito bom.

4.2 Comparação entre os processos de data mining tradicional e ILP

Técnica	Taxa de acerto	Precisão	Sensibilidade
Random Forest - 1000 árvores	68.00%	66.94%	70.81%
ILP	55.4%	67.8%	20.3%

Tabela 4.5: Comparação entre resultados Random Forest e ILP

A abordagem baseada em ILP teve resultados fracos comparados com os de data mining tradicional.Isto deve-se ao facto que no tempo projetado para esta dissertação não se conseguiu generalizar suficientes hipóteses para cobrir os exemplos fora do conjunto de treino. Para corrigir isto seria necessário estabelecer mais regras auxiliares como será referido na secção 5.2. Portanto, *Random Forest* de 1000 árvores continua a fornecer o melhor modelo de classificação. Considerou-se que a razão porque não se obteve melhores performances devido ao facto de o conjunto de dados não ser o suficientemente grande e ter qualidade suficiente (e.g: falta de dimensões mais relevantes) para inferir melhores modelos.

Capítulo 5

Conclusões e Trabalho Futuro

5.1 Conclusões

Esta dissertação foi realizada com o intuito de se construir um modelo de classificação capaz de prever casos de reinternamento em organizações hospitalares, e dar o alarme quando existe um internamento que seja um potencial caso de risco. Podemos considerar este objetivo atingido, após um processo de KDD sobre um conjunto de dados de um hospital, tendo sido criados vários modelos de classificação com base em *Naive Bayes*, *Random Forests*, SVMs e ILP. Entre estes, a *Random Forest* de 1000 árvores apresentou melhores resultados, tendo predito corretamente mais de 70% dos casos que originam reinternamentos com uma taxa de acerto de 68%. Estes resultados são considerados positivos e considera-se que o modelo satisfaz o objetivo da dissertação. É de notar que o modelo de ILP obteve resultados consideravelmente fracos, mas mantemo-nos otimistas que com mais trabalho e seguindo as recomendações propostas na secção de trabalho futuro haja potencial para se obter resultados muito melhores. Caso após essas recomendações, continuar-se a obter resultados fracos, têm que se concluir que ILP não é uma campo adequado para este tipo de problemas.

5.2 Trabalho Futuro

Uma das principais limitações deste projeto foi o limite de recursos computacionais, processos de data mining e ILP requerem dezenas de gigabytes de memoria RAM para o seu funcionamento. Maiores recursos computacionais permitem a implementação de técnicas mais complexas capazes de apresentar melhores resultados. Um campo não explorado, mas que pode ter interesse para este problema, é o das redes neuronais [KJ13] [GCF+12]. Estes são modelos computacionais que pretendem simular o funcionamento de um cérebro no aspeto em que cada neurónio existe numa rede de neurónios densamente interligada, e que estes se encontraram em constante comunicação uns com os outros. A força dos sinais de

cada neurónio adapta-se ao longo do tempo consoante os sinais recebidos dos vizinhos. A principais vantagens das redes neuronais em relação a outros métodos de data mining são a sua capacidade de processamento em paralelo, a sua tolerância a falhas e ruídos e a capacidade de encontrar soluções para problemas que não possuem um algoritmo ou outro método de resolução definido. A implementação da rede neuronal conhecida por *multi-layer perceptron* figurava-se inicialmente nos planos desta dissertação, mas foi abandonada em favor de se experimentar a abordagem de ILP.

No *multi-layer perceptron* cada nódulo irá realizar a soma do produto entre os vários inputs recebidos e o peso associado à ligação, essa soma irá passar por uma função de ativação não linear sendo o resultado deste cálculo propagado para a camada seguinte. Este processo será repetido em todas as camadas até chegar à camada de output, os nódulos desta camada final encontram-se associados a cada uma das classes presentes no conjunto de dados. Para a realização a aprendizagem MP usam uma técnica conhecida com *backpropergation*. A *backpropergation* está dividida em dois passos. O primeiro passo consiste na realização da método de propagação mencionado e a obtenção do valor de saída, este irá ser comparado com o resultado esperado e a diferença entre eles indica o valor de erro. No segundo passo o erro será propagado pela rede no sentido inverso de forma a os vários nódulos ajustarem os seus pesos com o objetivo de diminuir o erro nas próximas iterações. Através deste processo os *multi-layer perceptron* realizam a aprendizagem e criam modelos capazes de realizar classificações.

Também de notar que para a realização de modelo não estiveram disponíveis dados relativamente ao historial médico do paciente, na investigação do trabalho nesta área notou-se que estes dados encontram-se sempre presentes e acredita-se que com estes será possível criar modelos melhores preditivos.

A abordagem de ILP, sendo esta supervisionada, pode beneficiar do conhecimento de alguém familiar com a área de saúde para a criação de mais regras auxiliares, o que se supõe que terá uma significante influencia positiva nos resultados finais. Além disso, a questão dos recursos computacionais volta a ser relevante, várias regras mais complexas foram descartadas devido ao facto de aumentarem o tempos de execução a níveis demasiados elevadas. Dois exemplos desta situação foram os seguintes conjuntos de regras:

```
1
2 nomeciedosegteq(F,X,Y):-
3 consumo(_,_,_,_,_,F,_,X,_,_,_,_,_),number(X), number(Y),
4   X >= Y, !.
5
6 nomeciedoselteq(F,X,Y):-
7 consumo(_,_,_,_,_,F,_,X,_,_,_,_,_),number(X), number(Y),
8   X  =< Y, !.
9
```

```
10 nomecieqtadmgteq(F,X,Y):-
11 consumo(_,X,_,_,_,F,_,_,_,_,_,_,_),number(X), number(Y),
12   X >= Y, !.
13
14 nomecieqtadmlteq(F,X,Y):-
15 consumo(_,X,_,_,_,F,_,_,_,_,_,_,_),number(X), number(Y),
16   X  =< Y, !.
17
18 nomecieqtdiagteq(F,X,Y):-
19 consumo(_,_,X,_,_,F,_,_,_,_,_,_,_),number(X), number(Y),
20   X >= Y, !.
21
22 nomecieqtdialteq(F,X,Y):-
23 consumo(_,_,X,_,_,F,_,_,_,_,_,_,_),number(X), number(Y),
24   X  =< Y, !.
```

Este conjunto tem como objectivo permitir a criação de hipóteses $\leq$ e $\geq$ específicas para os consumos, tendo em conta que é necessário realizar uma distinção entre doses de medicamentos distintos.

```
1
2 conlteq(T,F):-consumo(_,_,_,_,_,_,T,_,_,_,_,_,_),count(consumo(_,_,_,_,_,_,T,
    _,_,_,_,_,_),F), number(Y),F =< Y, !.
3 congteq(T,F):-consumo(_,_,_,_,_,_,T,_,_,_,_,_,_),count(consumo(_,_,_,_,_,_,T,
    _,_,_,_,_,_),F), number(Y),F >= Y, !.
4 cirlteq(T,F):-consumo(F,_,_,_,_,_,_,_,_),count(consumo(T,_,_,_,_,_,_,_,_),F),
     number(Y),F =< Y, !.
5 cirgteq(T,F):-consumo(F,_,_,_,_,_,_,_,_),count(consumo(T,_,_,_,_,_,_,_,_),F),
    number(Y),F >= Y, !.
6 intlteq(T,F):-consumo(T,_,_,_,_,_),count(consumo(T,_,_,_,_,_),F),number(Y), F
     =< Y, !.
7 intgteq(T,F):-consumo(T,_,_,_,_,_),count(consumo(T,_,_,_,_,_),F),number(Y), F
     >= Y, !.
8
9 count(P,Count) :-
10        findall(1,P,L),
11        length(L,Count).
```

Estas regras têm como objectivo permitir a criação de hipóteses de $\leq$ e $\geq$ em relação à quantidade de consumos, de cirurgias de um episódio e de intervenções de uma cirurgia. Infelizmente, foi considerado que estes dois conjuntos de regras aumentam demasiado o tempo de execução e tiveram de ser descartadas. Portanto, melhores máquinas permitiram a execução de técnicas mais complexas e a remoção de alguns limites como a do *RecallNumber* para a tabela de consumos.

Finalmente, faltará ainda a integração do programa nas soluções da Glintt. Rapidminer possui

uma API que permite a fácil integração em java, esta pode ser usada em serviços .NET através do uso de frameworks como *Java native interface*, em alternativa é possível invocar o Rapidminer pela linha de comandos ou através de vários webservices fornecidos pela a empresa responsável pelo Rapidminer. Para a integração do modelo criado através de ILP, basta extrair as hipóteses da teoría e convertê-las para a linguagem desejada.

Referências

[BKNS00] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng e Jörg Sander. Lof: Identifying density-based local outliers. *SIGMOD Rec.*, 29(2):93–104, May 2000. URL: http://doi.acm.org/10.1145/335191.335388, doi:10.1145/335191.335388.

[BPO12] José Augusto Baranauskas, Pedro Santoro Perez e Thais Mayumi Oshiro. How many trees in a random forest? In *Machine Learning and Data Mining in Pattern Recognition*, volume 7376 of *Lecture Notes in Computer Science*. Springer, 2012.

[Bre99] Leo Breiman. Random forests-random features. Technical report, University of California Berkeley, September 1999. URL: http://www.stat.berkeley.edu/~breiman/random-forests.pdf.

[CL11] Chih-Chung Chang e Chih-Jen Lin. Libsvm - a library for support vector machines, May 2011.

[Con08] Joao Conceicão. The aleph system made easy. Technical report, FEUP, Julho 2008. URL: http://paginas.fe.up.pt/~ee03250/database/DissertationThesis-TheAlephSystemMadeEasy-FinalVersion.pdf.

[dis] Disadvantages of support vector machines. URL: http://www.svms.org/disadvantages.html.

[eSW13] James Natale e Shengyong Wange. A decision tree model for predicting heart failure patient readmissions. In *Proceedings of the 2013 Industrial and Systems Engineering Research Conference*, 2013. URL: https://www.academia.edu/3597479/A_Decision_Tree_Mode_for_Predicting_Heart_Failure_Readmissions.

[FCV+12] André S. Fialho, Federico Cismondi, Susana M. Vieira, Shane R. Reti, Joao M. C. Sousa e Stan N. Finkelstein. Data mining using clinical physiology at discharge to predict icu readmissions. *Expert Syst. Appl.*, 39(18):13158–13165, 2012. URL: http://dblp.uni-trier.de/db/journals/eswa/eswa39.html#FialhoCVRSF12.

[Fla10] Peter Flach. Roc analysis. pages 869–875, 2010.

[FPSS96] Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 1996.

[GCF+12] João Gama, André Carvalho, Katti Faceli, Ana Lorena e Márcia Oliveira. *Extração de Conhecimento de Dados - Data Mining*. Edições Silabo, R. Cidade de Manchester, Lisboa, 2012.

[Hal00] Mark H. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 359–366, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[HIV+13] Arian Hosseinzadeh, Masoumeh T. Izadi, Aman Verma, Doina Precup e David L. Buckeridge. Assessing the predictability of hospital readmission using machine learning. In *Proceedings of the Twenty-Fifth Innovative Applications of Artificial Intelligence Conference, IAAI 2013, July 14-18, 2013, Bellevue, Washington, USA.*, 2013.

[KJ13] Max Kuhn e Kjell Johnson. *Applied Predictive Modelling*. Springer, New York, 2013.

[LD94] Nada Lavrac e Saso Dazeroski. *Inductive Logic Programming - Techniques and Applications"*. Ellis Horwood, New York, 1994.

[MDYO04] Mohd Saberi Mohamad, Safaai Deris, Safie Mat Yatim e Muhammad Razib Othman. Feature selection method using genetic algorithm for the classifcation of small and high dimension data. In *First International Symposium on Information and Communications Technologies*, Putrajaya, Malaysia, 2004.

[Mov] Vida Movahedi. Conjunctive normal form & horn clauses. URL: http://www.eecs.yorku.ca/course_archive/2009-10/S/3401/slides/02_CNF_Horn.pdf.

[Mug95] S. Muggleton. Inverse Entailment and Progol. *New Generation Computing, Special issue on Inductive Logic Programming*, 13(3-4):245–286, 1995.

[Pia14] Gregory Piatetsky. Kdnuggets 15th annual analytics, data mining,-data science software poll: Rapidminer continues to lead, Junho 2014. URL: http://www.kdnuggets.com/2014/06/kdnuggets-annual-software-poll-rapidminer-continues-lead.html.

[RSK03] Marko Robnik-Sikonja e Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, pages 23–69, 2003.

[SB04] Carlos Soares e Pavel B. Brazdil. A meta-learning method to select the kernel width in support vector regression. *Machine Learning*, 54:195–209, 2004.

[Sri] Ashwin Srinivasan. The aleph manual. URL: http://www.cs.ox.ac.uk/activities/machlearn/Aleph/aleph.html.

[VAM+13] Kiyana Zolfagharand Nele Verbiest, Jayshree Agarwal, Naren Meadem, Si-Chi Chin, Senjuti Basu Roy, Ankur Teredesai, David Hazel, Paul Amoroso e Lester Reed. Predicting risk-of-readmission for congestive heart failure patients: A multi-layer approach. *CoRR*, abs/1306.2094, 2013. URL: http://arxiv.org/ftp/arxiv/papers/1306/1306.2094.pdf.

[VIM] Discussao: variable importance measure. URL: http://rapid-i.com/rapidforum/iindex.php?topic=417.0.

[web] Rapidminer:about us. URL: http://rapidminer.com/about-us/.

[WFH11] Ian H. Witten, Eibe Frank e Mark H. Hall. *Data Mining. Practical machine learning tools and techniques*. Morgan Kaufmann, 30 Corporate Drive, Suite 400, Burlington, MA 01803, USA, 2011.

[wHcCjL10] Chih wei Hsu, Chih chung Chang e Chih jen Lin. A practical guide to support vector classification, 2010.

REFERÊNCIAS

Anexo A

Processos de rapidminer

A.1 Seleção de dados

```
1
2  <?xml version="1.0" encoding="UTF-8" standalone="no"?>
3  <process version="5.3.013">
4    <context>
5      <input/>
6      <output/>
7      <macros/>
8    </context>
9    <operator activated="true" class="process" compatibility="5.3.013" expanded="true
       " name="Process">
10     <parameter key="logverbosity" value="all"/>
11     <parameter key="logfile" value="D:\testexrff.xrff"/>
12     <process expanded="true">
13       <operator activated="false" class="subprocess" compatibility="5.3.013"
           expanded="true" height="76" name="Data Selection" width="90" x="45" y="
           120">
14         <process expanded="true">
15           <operator activated="true" class="read_database" compatibility="5.3.013"
               expanded="true" height="60" name="Read Episodios" width="90" x="45" y
               ="120">
16             <parameter key="connection" value="DB"/>
17             <parameter key="query" value="&#10;&#10;&#10;Select R.*,D.*,convert(int
                 ,DATEDIFF(d, [DT_NASC], R.DT_INT)/365.25) AS IDADE FROM&#10;(select
                  *, 'S' as REINTERNAMENTO, DATEDIFF(d,a.[DT_INT],a.DT_ALTA) as
                 TEMPOINTER from [DWS_DM].[dbo].[sigt_ints_braga] a where exists (
                 select 0 from [DWS_DM].[dbo].[sigt_ints_braga]b &#10;where a.DOENTE
                  =b.DOENTE &#10;and a.EPISODIO!= b.EPISODIO&#10;and DATEDIFF(d,a.
                 DT_ALTA,b.[DT_INT]) between 0 and 5)&#10;&#10;UNION ALL&#10;&#10;
                 select *,'N' as REINTERNAMENTO,DATEDIFF(d,a.[DT_INT],a.DT_ALTA) as
```

```
                TEMPOINTER from [DWS_DM].[dbo].[sigt_ints_braga] a where NOT exists
                 (select 0 from [DWS_DM].[dbo].[sigt_ints_braga] b &#10;where a.
                DOENTE =b.DOENTE &#10;and a.EPISODIO!= b.EPISODIO&#10;and DATEDIFF(
                d,a.DT_ALTA,b.[DT_INT]) between 0 and 5))R,&#10;&#10;(SELECT [
                COD_DOENTE]&#10;&#9;  ,[COD_APLICACAO]&#10;&#9; ,[T_DOENTE]&#10;
                     ,[CENTRO_SAUDE]&#10;      ,[PAIS]&#10;      ,[DISTRITO]&#10;
                     ,[CONCELHO]&#10;      ,[FREGUESIA]&#10;      ,[LOCALIDADE
                ]&#10;      ,[PAIS_NASC]&#10;      ,[DISTRITO_NASC]&#10;      ,[
                CONCELHO_NASC]&#10;      ,[FREGUESIA_NASC]&#10;      ,[
                NACIONALIDADE]&#10;      ,[NATURALIDADE]&#10;      ,[DT_NASC]&#10;
                     ,[ESCOLARIDADE]&#10;      ,[ESTADO_CIVIL]&#10;      ,[
                ESTADO_PROFISSAO]&#10;      ,[PROFISSAO]&#10;      ,[ALTURA]&#10;
                     ,[FLAG_PROG_CTRL_HIV]&#10;      ,[GRUPO_SANG]&#10;      ,[
                MEDICO_FAMILIA]&#10;      ,[SEXO]&#10;    &#10;  FROM [DWS_DM].[dbo
                ].[DIM_DOENTE])D&#10;  WHERE  R.DOENTE= D.COD_DOENTE and R.T_DOENTE
                 =D.T_DOENTE and D.COD_APLICACAO ='GH';&#10;  "/>
18              <enumeration key="parameters"/>
19            </operator>
20            <operator activated="true" class="nominal_to_date" compatibility="5.3.013
                " expanded="true" height="76" name="Nominal to Date (3)" width="90" x
                ="180" y="120">
21              <parameter key="attribute_name" value="DT_NASC"/>
22              <parameter key="date_format" value="yyyy-MM-dd"/>
23            </operator>
24            <operator activated="true" class="date_to_numerical" compatibility="
                5.3.013" expanded="true" height="76" name="Date to Numerical (7)"
                width="90" x="313" y="120">
25              <parameter key="attribute_name" value="DT_NASC"/>
26              <parameter key="time_unit" value="year"/>
27              <parameter key="day_relative_to" value="year"/>
28              <parameter key="quarter_relative_to" value="epoch"/>
29            </operator>
30            <operator activated="true" class="numerical_to_polynominal" compatibility
                ="5.3.013" expanded="true" height="76" name="Numerical to Polynominal
                 (4)" width="90" x="450" y="120">
31              <parameter key="attribute_filter_type" value="single"/>
32              <parameter key="attribute" value="ID_DOENTE"/>
33            </operator>
34            <operator activated="true" class="read_database" compatibility="5.3.013"
                expanded="true" height="60" name="Read Cirurgias" width="90" x="45" y
                ="300">
35              <parameter key="connection" value="DB"/>
36              <parameter key="query" value=" SELECT R1.[EPISODIO],R1.
                    MEDIA_QTD_DURACAO_ANESTESIA,R1.MEDIA_QTD_DURACAO_CIRURGIA,R2.[
                    DESCRICAO_ANESTESIA],R2.TIPO_ASSEPSIA,R2.TIPO_ANESTESIA,R2.COD_ASA,
                    R2.TIPO_CIR&#10;  FROM&#10; (SELECT &#10;       [EPISODIO]&#10;
                         ,AVG(CONVERT(INT, [QTD_DURACAO_ANESTESIA]))AS
                    MEDIA_QTD_DURACAO_ANESTESIA&#10;      ,AVG(CONVERT(INT, [
                    QTD_DURACAO_CIRURGIA])) AS MEDIA_QTD_DURACAO_CIRURGIA&#10;  &#10;
```

```
                FROM [DWS_DM].[dbo].[sigt_cir_braga]&#10;  GROUP BY [EPISODIO]) R1
                ,&#10;  &#10; &#10; (Select T1.EPISODIO,T1.[DESCRICAO_ANESTESIA],T2
                .TIPO_ASSEPSIA,T3.TIPO_ANESTESIA,T4.COD_ASA,T5.TIPO_CIR from &#10;
                (SELECT DISTINCT&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.[
                DESCRICAO_ANESTESIA]&#10;&#9;&#9;FROM [DWS_DM].[dbo].[
                sigt_cir_braga] F1&#10;&#9;&#9;OUTER APPLY&#10;&#9;&#9;&#9;(
                SELECT TOP 1 r.[DESCRICAO_ANESTESIA] , (COUNT(*)) AS freq
                &#10;&#9;&#9;&#9;&#9;&#9;FROM  [DWS_DM].[dbo].[sigt_cir_braga] R
                &#10;&#9;&#9;&#9;&#9;&#9;WHERE r.EPISODIO = F1.EPISODIO
                &#10;&#9;&#9;&#9;&#9;&#9;GROUP BY [DESCRICAO_ANESTESIA]
                &#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count([DESCRICAO_ANESTESIA]) DESC
                &#10;&#9;&#9;&#9;) x)T1,&#10;&#9;(SELECT DISTINCT&#10;&#9;&#9;F1.
                EPISODIO,&#10;&#9;&#9;x.TIPO_ASSEPSIA&#10;&#9;&#9;FROM [DWS_DM].[
                dbo].[sigt_cir_braga] F1&#10;&#9;&#9;OUTER APPLY&#10;&#9;&#9;&#9;(
                      SELECT TOP 1 r.TIPO_ASSEPSIA , (COUNT(*)) AS freq
                &#10;&#9;&#9;&#9;&#9;&#9;FROM  [DWS_DM].[dbo].[sigt_cir_braga] R
                &#10;&#9;&#9;&#9;&#9;&#9;WHERE r.EPISODIO = F1.EPISODIO
                &#10;&#9;&#9;&#9;&#9;&#9;GROUP BY TIPO_ASSEPSIA
                &#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(TIPO_ASSEPSIA) DESC
                &#10;&#9;&#9;&#9;) x)T2,    &#10;
37  (SELECT DISTINCT&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.TIPO_ANESTESIA
    &#10;&#9;&#9;FROM [DWS_DM].[dbo].[sigt_cir_braga] F1&#10;&#9;&#9;OUTER APPLY
    &#10;&#9;&#9;&#9;(       SELECT TOP 1 r.TIPO_ANESTESIA , (COUNT(*)) AS freq
    &#10;&#9;&#9;&#9;&#9;&#9;FROM  [DWS_DM].[dbo].[sigt_cir_braga] R
    &#10;&#9;&#9;&#9;&#9;&#9;WHERE r.EPISODIO = F1.EPISODIO
    &#10;&#9;&#9;&#9;&#9;&#9;GROUP BY TIPO_ANESTESIA &#10;&#9;&#9;&#9;&#9;&#9;
    ORDER BY count(TIPO_ANESTESIA) DESC&#10;&#9;&#9;&#9;) x)T3,&#10;&#9;(SELECT
    DISTINCT&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.COD_ASA&#10;&#9;&#9;FROM [
    DWS_DM].[dbo].[sigt_cir_braga] F1&#10;&#9;&#9;OUTER APPLY&#10;&#9;&#9;&#9;(
          SELECT TOP 1 r.COD_ASA , (COUNT(*)) AS freq&#10;&#9;&#9;&#9;&#9;&#9;
    FROM  [DWS_DM].[dbo].[sigt_cir_braga] R&#10;&#9;&#9;&#9;&#9;&#9;WHERE r.
    EPISODIO = F1.EPISODIO&#10;&#9;&#9;&#9;&#9;&#9;GROUP BY COD_ASA
    &#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(COD_ASA) DESC&#10;&#9;&#9;&#9;) x)T4
    ,&#10;&#9;&#9;&#9; (SELECT DISTINCT&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.[
    TIPO_CIR]&#10;&#9;&#9;FROM [DWS_DM].[dbo].[sigt_cir_braga] F1&#10;&#9;&#9;
    OUTER APPLY&#10;&#9;&#9;&#9;(       SELECT TOP 1 r.[TIPO_CIR] , (COUNT(*))
    AS freq&#10;&#9;&#9;&#9;&#9;&#9;FROM  [DWS_DM].[dbo].[sigt_cir_braga] R
    &#10;&#9;&#9;&#9;&#9;&#9;WHERE r.EPISODIO = F1.EPISODIO
    &#10;&#9;&#9;&#9;&#9;&#9;GROUP BY [TIPO_CIR] &#10;&#9;&#9;&#9;&#9;&#9;ORDER
    BY count([TIPO_CIR]) DESC&#10;&#9;&#9;&#9;) x)T5&#10;    &#10;    WHERE T1.[
    EPISODIO] = T2.[EPISODIO] AND T1.[EPISODIO] = T3.[EPISODIO] AND T1.[EPISODIO
    ] = T4.[EPISODIO] AND T1.[EPISODIO] = T5.[EPISODIO]) R2&#10;    WHERE R1.
    EPISODIO= R2.EPISODIO&#10;    "/>
38          <enumeration key="parameters"/>
39        </operator>
40        <operator activated="true" class="read_database" compatibility="5.3.013"
            expanded="true" height="60" name="Read Intervencoes" width="90" x="45
            " y="210">
41          <parameter key="connection" value="DB"/>
```

```
42            <parameter key="query" value="&#10;SELECT T1.[EPISODIO],T1.[N_REG_OPER
                  ],T1.[DESCRICAO] AS INTERV_DESC_PRINCIPAL,T1.[DT_OPER]AS
                  INTERV_DT_OPER_PRINCIPAL, T2.[DESCRICAO] AS INTERV_DESC_SEC1 ,T2.[
                  DT_OPER] AS INTERV_DT_OPER_SEC1,T3.[DESCRICAO] AS INTERV_DESC_SEC2
                  ,T3.[DT_OPER] AS INTERV_DT_OPER_SEC2, T4.NR_INTER, T6.NR_CIR AS
                  NR_CIR FROM &#9;&#10;&#9;(SELECT [EPISODIO],  &#10;&#9;&#9;   [
                  N_REG_OPER]&#10;&#9;&#9;  ,[DESCRICAO],[FLAG_PRINCIPAL],[N_ORD],[
                  DT_OPER]&#10;&#9;  FROM [DWS_DM].[dbo].[sigt_interv_braga] Where
                  FLAG_PRINCIPAL ='P')T1&#10;&#9;  LEFT JOIN &#10;&#9;  (SELECT
                  &#10;&#9;&#9;   [N_REG_OPER]&#10;&#9;&#9;  ,[DESCRICAO],[
                  FLAG_PRINCIPAL],[N_ORD],[DT_OPER]&#10;&#9;  FROM [DWS_DM].[dbo].[
                  sigt_interv_braga] Where N_ORD =2) T2&#10;&#9;  ON T1.N_REG_OPER =
                  T2.N_REG_OPER  &#10;&#9;  &#10;&#9;  LEFT JOIN &#10;&#9;  (SELECT
                  &#10;&#9;&#9;   [N_REG_OPER]&#10;&#9;&#9;  ,[DESCRICAO],[
                  FLAG_PRINCIPAL],[N_ORD],[DT_OPER]&#10;&#9;  FROM [DWS_DM].[dbo].[
                  sigt_interv_braga]  Where N_ORD =3)T3 ON&#10;&#9;  T1.N_REG_OPER =
                  T3.N_REG_OPER, &#10;&#9;  (SELECT T62.EPISODIO, SUM (T61.NR_INTER)
                  AS NR_INTER&#10;FROM (&#10;SELECT  N_REG_OPER, COUNT (*)AS NR_INTER
                   FROM [DWS_DM].[dbo].[sigt_interv_braga] group by N_REG_OPER) T61
                  ,&#10;[DWS_DM].[dbo].[sigt_cir_braga] T62&#10;WHERE T62.N_REG_OPER
                  = T61.N_REG_OPER GROUP BY EPISODIO )T4,&#10;&#9;  (SELECT
                  &#10;&#9;&#9;&#9;[EPISODIO], MIN([N_REG_OPER]) as [N_REG_OPER
                  ]&#10;&#9;&#9;FROM&#10;&#9;&#9;&#9;[DWS_DM].[dbo].[
                  sigt_interv_braga]&#10;&#9;&#9;GROUP BY&#10;&#9;&#9;&#9;[EPISODIO])
                  T5,&#10;&#9;&#9;(SELECT EPISODIO , COUNT (*) as NR_CIR FROM [DWS_DM
                  ].[dbo].[sigt_cir_braga]Group by EPISODIO)T6&#10;&#9;  WHERE  T4.
                  EPISODIO=T1.EPISODIO and T5.[N_REG_OPER]=T1.[N_REG_OPER]AND T6.
                  EPISODIO = T1.EPISODIO "/>
43            <enumeration key="parameters"/>
44          </operator>
45          <operator activated="true" class="read_database" compatibility="5.3.013"
                expanded="true" height="60" name="Read Consumos" width="90" x="45" y=
                "30">
46            <parameter key="connection" value="DB"/>
47            <parameter key="query" value="SELECT R1.*, R2.
                  MEDIA_DOSE_PRINCIPIO_ACTIVO,R2.MEDIA_QT_ADM_PRINCIPIO_ACTIVO,R2.
                  MEDIA_QT_DIA_PRINCIPIO_ACTIVO,R3.MEDIA_DOSE_NOME_CIENT,R3.
                  MEDIA_QT_DIA_NOME_CIENT,R3.MEDIA_QT_ADM_NOME_CIENT FROM&#10;  &#10;
                   (SELECT T1.EPISODIO, T1.NOME_CIENT as NOME_CIENT_FREQ,T2.
                  PRINCIPIO_ACTIVO as PRINCIPIO_ACTIVO_FREQ,T3.NOME_COMERCIAL AS
                  NOME_COMERCIAL_FREQ, T4.VIA_ADM AS VIA_ADM_FREQ,T5.HORARIO AS
                  HORARIO_FREQ ,T6.FREQ AS ADM_FREQ, T7.PRESC_POS,T8.PRESC_NEG&#10;
                  FROM&#10; &#10; (SELECT DISTINCT&#10;&#9;&#9;F1.EPISODIO
                  ,&#10;&#9;&#9;x.NOME_CIENT&#10;&#9;&#9;FROM [DWS_DM].[dbo].[
                  sigt_consumos_braga] F1&#10;&#9;&#9;OUTER APPLY&#10;&#9;&#9;(
                        SELECT TOP 1 r.NOME_CIENT , (COUNT(*)) AS freq
                  &#10;&#9;&#9;&#9;&#9;FROM  [DWS_DM].[dbo].[sigt_consumos_braga]
                   R&#10;&#9;&#9;&#9;&#9;WHERE r.EPISODIO = F1.EPISODIO
                  &#10;&#9;&#9;&#9;&#9;GROUP BY NOME_CIENT
```

```
&#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(NOME_CIENT) DESC
&#10;&#9;&#9;) x )T1,&#10;&#9;&#9;&#9;&#10;( SELECT DISTINCT
&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.PRINCIPIO_ACTIVO
&#10;&#9;&#9;FROM [DWS_DM].[dbo].[sigt_consumos_braga] F1
&#10;&#9;&#9;OUTER APPLY&#10;&#9;&#9;&#9;(       SELECT TOP 1 r.
PRINCIPIO_ACTIVO , (COUNT(*)) AS freq&#10;&#9;&#9;&#9;&#9;&#9;FROM
 [DWS_DM].[dbo].[sigt_consumos_braga] R&#10;&#9;&#9;&#9;&#9;&#9;&#9;
WHERE r.EPISODIO = F1.EPISODIO&#10;&#9;&#9;&#9;&#9;&#9;GROUP BY
PRINCIPIO_ACTIVO &#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(
PRINCIPIO_ACTIVO) DESC&#10;&#9;&#9;&#9;) x )T2,&#10;
(SELECT DISTINCT&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.NOME_COMERCIAL
&#10;&#9;&#9;FROM [DWS_DM].[dbo].[sigt_consumos_braga] F1
&#10;&#9;&#9;OUTER APPLY&#10;&#9;&#9;&#9;(       SELECT TOP 1 r.
NOME_COMERCIAL , (COUNT(*)) AS freq&#10;&#9;&#9;&#9;&#9;&#9;FROM  [
DWS_DM].[dbo].[sigt_consumos_braga] R&#10;&#9;&#9;&#9;&#9;&#9;WHERE
 r.EPISODIO = F1.EPISODIO&#10;&#9;&#9;&#9;&#9;&#9;GROUP BY
NOME_COMERCIAL &#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(
NOME_COMERCIAL) DESC&#10;&#9;&#9;&#9;) x )T3,&#10;&#9;&#9;&#9;(
SELECT DISTINCT&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.VIA_ADM
&#10;&#9;&#9;FROM [DWS_DM].[dbo].[sigt_consumos_braga] F1
&#10;&#9;&#9;OUTER APPLY&#10;&#9;&#9;&#9;(       SELECT TOP 1 r.
VIA_ADM , (COUNT(*)) AS freq&#10;&#9;&#9;&#9;&#9;&#9;FROM  [DWS_DM
].[dbo].[sigt_consumos_braga] R&#10;&#9;&#9;&#9;&#9;&#9;WHERE r.
EPISODIO = F1.EPISODIO&#10;&#9;&#9;&#9;&#9;&#9;GROUP BY VIA_ADM
&#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(VIA_ADM) DESC
&#10;&#9;&#9;&#9;) x )T4,&#10;&#9;&#9;&#9;(SELECT DISTINCT
&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.HORARIO&#10;&#9;&#9;FROM [
DWS_DM].[dbo].[sigt_consumos_braga] F1&#10;&#9;&#9;OUTER APPLY
&#10;&#9;&#9;&#9;(       SELECT TOP 1 r.HORARIO , (COUNT(*)) AS
freq&#10;&#9;&#9;&#9;&#9;&#9;FROM  [DWS_DM].[dbo].[
sigt_consumos_braga] R&#10;&#9;&#9;&#9;&#9;&#9;WHERE r.EPISODIO =
F1.EPISODIO&#10;&#9;&#9;&#9;&#9;GROUP BY HORARIO
&#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(HORARIO) DESC
&#10;&#9;&#9;&#9;) x )T5,&#10;&#9;&#9;(SELECT DISTINCT
&#10;&#9;&#9;F1.EPISODIO,&#10;&#9;&#9;x.freq&#10;&#9;&#9;FROM [
DWS_DM].[dbo].[sigt_consumos_braga] F1&#10;&#9;&#9;OUTER APPLY
&#10;&#9;&#9;&#9;(       SELECT TOP 1 r.freq ,
(COUNT(*)) AS freq1&#10;&#9;&#9;&#9;&#9;&#9;FROM  [DWS_DM].[dbo].[
sigt_consumos_braga] R&#10;&#9;&#9;&#9;&#9;&#9;WHERE r.EPISODIO =
F1.EPISODIO&#10;&#9;&#9;&#9;&#9;GROUP BY freq
&#10;&#9;&#9;&#9;&#9;&#9;ORDER BY count(freq) DESC
&#10;&#9;&#9;&#9;) x )T6,&#10;&#9;&#9;&#9;  (SELECT
&#10;&#9;&#9;&#9;&#9;&#9;EPISODIO,COUNT(CASE WHEN PRESCRITO = 'S'
THEN 1 END) AS PRESC_POS&#10;&#9;&#9;&#9;&#9;FROM
&#10;&#9;&#9;&#9;&#9;   [DWS_DM].[dbo].[sigt_consumos_braga
]&#10;&#9;&#9;&#9;&#9;GROUP BY EPISODIO)T7,&#10;&#9;&#9;&#9;(SELECT
&#10;&#9;&#9;&#9;&#9;&#9;EPISODIO,COUNT(CASE WHEN PRESCRITO = 'N'
THEN 1 END) AS PRESC_NEG&#10;&#9;&#9;&#9;&#9;FROM
&#10;&#9;&#9;&#9;&#9;   [DWS_DM].[dbo].[sigt_consumos_braga
```

```
                ]&#10;&#9;&#9;&#9;&#9;GROUP BY EPISODIO)T8&#10;&#9;&#9;&#9;&#9;
                WHERE &#10;&#9;T1.EPISODIO= T2.EPISODIO AND &#10;&#9;T2.EPISODIO =
                T3.EPISODIO AND&#10;&#9;T4.EPISODIO =T3.EPISODIO AND &#10;&#9;T5.
                EPISODIO =T3.EPISODIO AND&#10;&#9;T6.EPISODIO =T3.EPISODIO AND
                &#10;&#9;T6.EPISODIO =T7.EPISODIO AND &#10;&#9;T8.EPISODIO =T3.
                EPISODIO
                &#10;&#9;&#9;&#9;&#9;&#10;&#9;&#9;&#9;&#9;&#10;&#9;&#9;&#9;&#9;&#10
50 ;&#9;&#9;&#9;&#9;)R1 LEFT JOIN&#10;&#9;&#9;&#9;&#9;&#10;&#9;&#10;   &#9;(SELECT
    &#10;&#9;&#9;&#9;&#9;[EPISODIO],PRINCIPIO_ACTIVO&#10;&#9;&#9;&#9;&#9;,AVG(
    CONVERT(float, dose))AS MEDIA_DOSE_PRINCIPIO_ACTIVO&#10;&#9;&#9;&#9;&#9;,AVG(
    CONVERT(float, QT_DIA))AS MEDIA_QT_DIA_PRINCIPIO_ACTIVO&#10;&#9;&#9;&#9;&#9;,
    AVG(CONVERT(float, QT_ADM)) AS MEDIA_QT_ADM_PRINCIPIO_ACTIVO
    &#10;&#9;&#9;&#9;&#9;FROM [DWS_DM].[dbo].[sigt_consumos_braga
    ]&#10;&#9;&#9;&#9;&#9;GROUP BY [EPISODIO],PRINCIPIO_ACTIVO&#10;&#9;  )
51              R2 ON R1.EPISODIO =R2.EPISODIO AND R1.PRINCIPIO_ACTIVO_FREQ=R2.
                    PRINCIPIO_ACTIVO LEFT JOIN&#10;&#9;   (SELECT &#10;&#9;&#9;   [
                    EPISODIO],NOME_CIENT&#10;&#9;&#9;  ,AVG(CONVERT(float, dose))AS
                    MEDIA_DOSE_NOME_CIENT&#10;&#9;&#9;  ,AVG(CONVERT(float, QT_DIA))AS
                    MEDIA_QT_DIA_NOME_CIENT&#10;&#9;&#9;  ,AVG(CONVERT(float, QT_ADM))
                    AS MEDIA_QT_ADM_NOME_CIENT&#10;&#9;&#9;   FROM [DWS_DM].[dbo].[
                    sigt_consumos_braga]&#10;&#9;&#9;   GROUP BY [EPISODIO],NOME_CIENT
                    &#10;&#9;  )R3 ON &#10;&#9;&#10;&#9;R1.EPISODIO =R3.EPISODIO AND R1
                    .NOME_CIENT_FREQ =R3.NOME_CIENT"/>
52              <enumeration key="parameters"/>
53            </operator>
54            <operator activated="true" class="nominal_to_date" compatibility="5.3.013
                  " expanded="true" height="76" name="Nominal to Date (8)" width="90" x
                  ="179" y="210">
55              <parameter key="attribute_name" value="INTERV_DT_OPER_SEC1"/>
56              <parameter key="date_format" value="yyyy-MM-dd"/>
57            </operator>
58            <operator activated="true" class="nominal_to_date" compatibility="5.3.013
                  " expanded="true" height="76" name="Nominal to Date (9)" width="90" x
                  ="313" y="210">
59              <parameter key="attribute_name" value="INTERV_DT_OPER_PRINCIPAL"/>
60              <parameter key="date_format" value="yyyy-MM-dd"/>
61            </operator>
62            <operator activated="true" class="nominal_to_date" compatibility="5.3.013
                  " expanded="true" height="76" name="Nominal to Date (10)" width="90"
                  x="447" y="210">
63              <parameter key="attribute_name" value="INTERV_DT_OPER_SEC2"/>
64              <parameter key="date_format" value="yyyy-MM-dd"/>
65            </operator>
66            <operator activated="true" class="date_to_numerical" compatibility="
                  5.3.013" expanded="true" height="76" name="Date to Numerical (8)"
                  width="90" x="581" y="210">
67              <parameter key="attribute_name" value="INTERV_DT_OPER_PRINCIPAL"/>
68              <parameter key="time_unit" value="day"/>
69              <parameter key="day_relative_to" value="year"/>
```

```
70            <parameter key="quarter_relative_to" value="epoch"/>
71          </operator>
72          <operator activated="true" class="date_to_numerical" compatibility="
                5.3.013" expanded="true" height="76" name="Date to Numerical (9)"
                width="90" x="715" y="210">
73            <parameter key="attribute_name" value="INTERV_DT_OPER_SEC1"/>
74            <parameter key="time_unit" value="day"/>
75            <parameter key="day_relative_to" value="year"/>
76            <parameter key="quarter_relative_to" value="epoch"/>
77          </operator>
78          <operator activated="true" class="date_to_numerical" compatibility="
                5.3.013" expanded="true" height="76" name="Date to Numerical (10)"
                width="90" x="849" y="210">
79            <parameter key="attribute_name" value="INTERV_DT_OPER_SEC2"/>
80            <parameter key="time_unit" value="day"/>
81            <parameter key="day_relative_to" value="year"/>
82            <parameter key="quarter_relative_to" value="epoch"/>
83          </operator>
84          <operator activated="true" class="remove_duplicates" compatibility="
                5.3.013" expanded="true" height="76" name="Remove Duplicates (2)"
                width="90" x="983" y="210">
85            <parameter key="attribute_filter_type" value="single"/>
86            <parameter key="attribute" value="N_REG_OPER"/>
87          </operator>
88          <operator activated="true" class="join" compatibility="5.3.013" expanded=
                "true" height="76" name="Join 1" width="90" x="1117" y="255">
89            <parameter key="use_id_attribute_as_key" value="false"/>
90            <list key="key_attributes">
91              <parameter key="EPISODIO" value="EPISODIO"/>
92            </list>
93          </operator>
94          <operator activated="true" class="remove_duplicates" compatibility="
                5.3.013" expanded="true" height="76" name="Remove Duplicates (3)"
                width="90" x="1251" y="255">
95            <parameter key="attribute_filter_type" value="single"/>
96            <parameter key="attribute" value="EPISODIO"/>
97          </operator>
98          <operator activated="true" class="join" compatibility="5.3.013" expanded=
                "true" height="76" name="Join 2" width="90" x="1318" y="165">
99            <parameter key="join_type" value="left"/>
100           <parameter key="use_id_attribute_as_key" value="false"/>
101           <list key="key_attributes">
102             <parameter key="EPISODIO" value="EPISODIO"/>
103           </list>
104         </operator>
105         <operator activated="true" class="join" compatibility="5.3.013" expanded=
                "true" height="76" name="Join 3" width="90" x="1318" y="30">
106           <parameter key="use_id_attribute_as_key" value="false"/>
107           <list key="key_attributes">
```

```
108                 <parameter key="EPISODIO" value="EPISODIO"/>
109               </list>
110             </operator>
111             <operator activated="true" class="replace_missing_values" compatibility="
                    5.3.013" expanded="true" height="94" name="Replace Missing Values -
                    zeroes" width="90" x="1452" y="30">
112               <parameter key="attribute_filter_type" value="regular_expression"/>
113               <parameter key="regular_expression" value="QT.*|MEDIA.*|DT_NASC"/>
114               <parameter key="default" value="zero"/>
115               <list key="columns"/>
116             </operator>
117             <operator activated="true" class="replace_missing_values" compatibility="
                    5.3.013" expanded="true" height="94" name="Replace Missing Values -
                    Desconhecido" width="90" x="1452" y="165">
118               <parameter key="attribute_filter_type" value="subset"/>
119               <parameter key="attributes" value="ADM_FREQ|CENTRO_SAUDE|COD_ASA|
                        COD_DOENTE|CONCELHO|CONCELHO_NASC|DESCRICAO_ANESTESIA|DISTRITO|
                        DISTRITO_NASC|DT_ALTA|DT_INT|EPISODIO|ESCOLARIDADE|ESTADO_CIVIL|
                        ESTADO_PROFISSAO|FLAG_PROG_CTRL_HIV|FREGUESIA|FREGUESIA_NASC|
                        GRUPO_SANG|HORARIO_FREQ|INTERV_DESC_PRINCIPAL|INTERV_DESC_SEC1|
                        INTERV_DESC_SEC2|LOCALIDADE|MEDICO_FAMILIA|NACIONALIDADE|
                        NATURALIDADE|NOME_CIENT_FREQ|NOME_COMERCIAL_FREQ|N_REG_OPER|PAIS|
                        PAIS_NASC|PRINCIPIO_ACTIVO_FREQ|PROFISSAO|REINTERNAMENTO|SEXO|
                        TIPO_ANESTESIA|TIPO_ASSEPSIA|T_DOENTE|T_EPISODIO|VIA_ADM_FREQ||
                        FLAG_N_SNS|DATA_ADMISSAO|TIPO_CIR"/>
120               <parameter key="default" value="value"/>
121               <list key="columns"/>
122               <parameter key="replenishment_value" value="DESCONHECIDO"/>
123             </operator>
124             <operator activated="true" class="replace_missing_values" compatibility="
                    5.3.013" expanded="true" height="94" name="Replace Missing Values
                    (16)" width="90" x="1586" y="165">
125               <parameter key="attribute_filter_type" value="subset"/>
126               <parameter key="attributes" value="NR_CIR|NR_INTER|TempoInter||
                        INTERV_DT_OPER_SEC2|INTERV_DT_OPER_SEC1|INTERV_DT_OPER_PRINCIPAL"/>
127               <parameter key="default" value="value"/>
128               <list key="columns"/>
129               <parameter key="replenishment_value" value="-1"/>
130             </operator>
131             <operator activated="true" class="replace_missing_values" compatibility="
                    5.3.013" expanded="true" height="94" name="Replace Missing Values
                    (17)" width="90" x="1720" y="30">
132               <parameter key="attribute_filter_type" value="single"/>
133               <parameter key="attribute" value="ALTURA"/>
134               <list key="columns"/>
135             </operator>
136             <operator activated="true" class="discretize_by_bins" compatibility="
                    5.3.013" expanded="true" height="94" name="Discretize Altura" width="
                    90" x="1720" y="165">
```

```
137             <parameter key="attribute_filter_type" value="single"/>
138             <parameter key="attribute" value="ALTURA"/>
139             <parameter key="number_of_bins" value="20"/>
140           </operator>
141           <operator activated="true" class="rename_by_replacing" compatibility="
                  5.3.013" expanded="true" height="76" name="Remove all non-word char"
                  width="90" x="1854" y="30"/>
142           <operator activated="true" class="set_role" compatibility="5.3.013"
                  expanded="true" height="76" name="Set Role" width="90" x="1854" y="
                  165">
143             <parameter key="attribute_name" value="REINTERNAMENTO"/>
144             <parameter key="target_role" value="label"/>
145             <list key="set_additional_roles"/>
146           </operator>
147           <operator activated="true" class="subprocess" compatibility="5.3.013"
                  expanded="true" height="76" name="Sampling" width="90" x="1988" y="30
                  ">
148             <process expanded="true">
149               <operator activated="true" class="multiply" compatibility="5.3.013"
                      expanded="true" height="94" name="Multiply" width="90" x="45" y="
                      30"/>
150               <operator activated="true" class="filter_examples" compatibility="
                      5.3.013" expanded="true" height="76" name="Filter Reinternamentos
                       Positives" width="90" x="180" y="30">
151                 <parameter key="condition_class" value="attribute_value_filter"/>
152                 <parameter key="parameter_string" value="REINTERNAMENTO =S"/>
153               </operator>
154               <operator activated="true" class="sample" compatibility="5.3.013"
                      expanded="true" height="76" name="Sample Positives" width="90" x=
                      "315" y="30">
155                 <parameter key="sample_size" value="2027"/>
156                 <list key="sample_size_per_class"/>
157                 <list key="sample_ratio_per_class"/>
158                 <list key="sample_probability_per_class"/>
159               </operator>
160               <operator activated="true" class="filter_examples" compatibility="
                      5.3.013" expanded="true" height="76" name="Filter Reinternamentos
                       Negatives" width="90" x="179" y="210">
161                 <parameter key="condition_class" value="attribute_value_filter"/>
162                 <parameter key="parameter_string" value="REINTERNAMENTO=N"/>
163               </operator>
164               <operator activated="true" class="sample" compatibility="5.3.013"
                      expanded="true" height="76" name="Sample Negatvies" width="90" x=
                      "313" y="210">
165                 <parameter key="sample_size" value="2027"/>
166                 <list key="sample_size_per_class"/>
167                 <list key="sample_ratio_per_class"/>
168                 <list key="sample_probability_per_class"/>
169               </operator>
```

```
170             <operator activated="true" class="append" compatibility="5.3.013"
                    expanded="true" height="112" name="Append" width="90" x="581" y="
                    75"/>
171             <connect from_port="in 1" to_op="Multiply" to_port="input"/>
172             <connect from_op="Multiply" from_port="output 1" to_op="Filter
                    Reinternamentos Positives" to_port="example set input"/>
173             <connect from_op="Multiply" from_port="output 2" to_op="Filter
                    Reinternamentos Negatives" to_port="example set input"/>
174             <connect from_op="Filter Reinternamentos Positives" from_port="
                    example set output" to_op="Sample Positives" to_port="example set
                     input"/>
175             <connect from_op="Sample Positives" from_port="example set output"
                    to_op="Append" to_port="example set 1"/>
176             <connect from_op="Filter Reinternamentos Negatives" from_port="
                    example set output" to_op="Sample Negatvies" to_port="example set
                     input"/>
177             <connect from_op="Sample Negatvies" from_port="example set output"
                    to_op="Append" to_port="example set 2"/>
178             <connect from_op="Append" from_port="merged set" to_port="out 1"/>
179             <portSpacing port="source_in 1" spacing="0"/>
180             <portSpacing port="source_in 2" spacing="0"/>
181             <portSpacing port="sink_out 1" spacing="0"/>
182             <portSpacing port="sink_out 2" spacing="0"/>
183           </process>
184         </operator>
185         <operator activated="true" class="select_attributes" compatibility="
                5.3.013" expanded="true" height="76" name="Select Attributes (7)"
                width="90" x="2189" y="30">
186           <parameter key="attribute_filter_type" value="single"/>
187           <parameter key="attribute" value="T_DOENTE_1"/>
188           <parameter key="attributes" value="|EPISODIO|COD_DOENTE"/>
189           <parameter key="invert_selection" value="true"/>
190         </operator>
191         <connect from_op="Read Episodios" from_port="output" to_op="Nominal to
                Date (3)" to_port="example set input"/>
192         <connect from_op="Nominal to Date (3)" from_port="example set output"
                to_op="Date to Numerical (7)" to_port="example set input"/>
193         <connect from_op="Date to Numerical (7)" from_port="example set output"
                to_op="Numerical to Polynominal (4)" to_port="example set input"/>
194         <connect from_op="Numerical to Polynominal (4)" from_port="example set
                output" to_op="Join 2" to_port="left"/>
195         <connect from_op="Read Cirurgias" from_port="output" to_op="Join 1"
                to_port="right"/>
196         <connect from_op="Read Intervencoes" from_port="output" to_op="Nominal to
                 Date (8)" to_port="example set input"/>
197         <connect from_op="Read Consumos" from_port="output" to_op="Join 3"
                to_port="left"/>
198         <connect from_op="Nominal to Date (8)" from_port="example set output"
                to_op="Nominal to Date (9)" to_port="example set input"/>
```

```
199           <connect from_op="Nominal to Date (9)" from_port="example set output"
                  to_op="Nominal to Date (10)" to_port="example set input"/>
200           <connect from_op="Nominal to Date (10)" from_port="example set output"
                  to_op="Date to Numerical (8)" to_port="example set input"/>
201           <connect from_op="Date to Numerical (8)" from_port="example set output"
                  to_op="Date to Numerical (9)" to_port="example set input"/>
202           <connect from_op="Date to Numerical (9)" from_port="example set output"
                  to_op="Date to Numerical (10)" to_port="example set input"/>
203           <connect from_op="Date to Numerical (10)" from_port="example set output"
                  to_op="Remove Duplicates (2)" to_port="example set input"/>
204           <connect from_op="Remove Duplicates (2)" from_port="example set output"
                  to_op="Join 1" to_port="left"/>
205           <connect from_op="Join 1" from_port="join" to_op="Remove Duplicates (3)"
                  to_port="example set input"/>
206           <connect from_op="Remove Duplicates (3)" from_port="example set output"
                  to_op="Join 2" to_port="right"/>
207           <connect from_op="Join 2" from_port="join" to_op="Join 3" to_port="right"
                  />
208           <connect from_op="Join 3" from_port="join" to_op="Replace Missing Values
                  - zeroes" to_port="example set input"/>
209           <connect from_op="Replace Missing Values - zeroes" from_port="example set
                   output" to_op="Replace Missing Values - Desconhecido" to_port="
                  example set input"/>
210           <connect from_op="Replace Missing Values - Desconhecido" from_port="
                  example set output" to_op="Replace Missing Values (16)" to_port="
                  example set input"/>
211           <connect from_op="Replace Missing Values (16)" from_port="example set
                  output" to_op="Replace Missing Values (17)" to_port="example set
                  input"/>
212           <connect from_op="Replace Missing Values (17)" from_port="example set
                  output" to_op="Discretize Altura" to_port="example set input"/>
213           <connect from_op="Discretize Altura" from_port="example set output" to_op
                  ="Remove all non-word char" to_port="example set input"/>
214           <connect from_op="Remove all non-word char" from_port="example set output
                  " to_op="Set Role" to_port="example set input"/>
215           <connect from_op="Set Role" from_port="example set output" to_op="
                  Sampling" to_port="in 1"/>
216           <connect from_op="Sampling" from_port="out 1" to_op="Select Attributes
                  (7)" to_port="example set input"/>
217           <connect from_op="Select Attributes (7)" from_port="example set output"
                  to_port="out 1"/>
218           <portSpacing port="source_in 1" spacing="0"/>
219           <portSpacing port="sink_out 1" spacing="0"/>
220           <portSpacing port="sink_out 2" spacing="0"/>
221         </process>
222       </operator>
223       <portSpacing port="source_input 1" spacing="0"/>
224       <portSpacing port="sink_result 1" spacing="0"/>
225     </process>
```

```
  </operator>
</process>
```

A.2 Deteção de Outliers

```

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.3.013">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.3.013" expanded="true
      " name="Process">
    <parameter key="logverbosity" value="all"/>
    <parameter key="logfile" value="D:\testexrff.xrff"/>
    <process expanded="true">
      <operator activated="true" class="detect_outlier_lof" compatibility="5.3.013"
           expanded="true" height="76" name="Detect Outlier (LOF)" width="90" x="
          179" y="255"/>
      <operator activated="true" class="filter_examples" compatibility="5.3.013"
          expanded="true" height="76" name="Filter Examples" width="90" x="313" y="
          255">
        <parameter key="condition_class" value="attribute_value_filter"/>
        <parameter key="parameter_string" value="outlier&lt;2.0"/>
      </operator>
      <connect from_op="Detect Outlier (LOF)" from_port="example set output" to_op=
          "Filter Examples" to_port="example set input"/>
      <portSpacing port="source_input 1" spacing="0"/>
      <portSpacing port="sink_result 1" spacing="0"/>
    </process>
  </operator>
</process>
```

A.3 Filtro

```

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.3.013">
  <context>
```

```
5     <input/>
6     <output/>
7     <macros/>
8   </context>
9   <operator activated="true" class="process" compatibility="5.3.013" expanded="true
        " name="Process">
10    <parameter key="logverbosity" value="all"/>
11    <parameter key="logfile" value="D:\testexrff.xrff"/>
12    <process expanded="true">
13      <operator activated="true" class="subprocess" compatibility="5.3.013"
            expanded="true" height="94" name="Preprocessing (2)" width="90" x="447" y
            ="255">
14        <process expanded="true">
15          <operator activated="true" class="weka:W-ReliefFAttributeEval"
                compatibility="5.3.000" expanded="true" height="76" name="W-
                ReliefFAttributeEval (2)" width="90" x="447" y="165"/>
16          <operator activated="true" class="select_by_weights" compatibility="
                5.3.013" expanded="true" height="94" name="Select by Weights (2)"
                width="90" x="581" y="165">
17            <parameter key="weight" value="0.01"/>
18          </operator>
19          <connect from_port="in 1" to_op="W-ReliefFAttributeEval (2)" to_port="
                example set"/>
20          <connect from_op="W-ReliefFAttributeEval (2)" from_port="weights" to_op="
                Select by Weights (2)" to_port="weights"/>
21          <connect from_op="W-ReliefFAttributeEval (2)" from_port="example set"
                to_op="Select by Weights (2)" to_port="example set input"/>
22          <connect from_op="Select by Weights (2)" from_port="example set output"
                to_port="out 1"/>
23          <connect from_op="Select by Weights (2)" from_port="weights" to_port="out
                 2"/>
24          <portSpacing port="source_in 1" spacing="0"/>
25          <portSpacing port="source_in 2" spacing="0"/>
26          <portSpacing port="sink_out 1" spacing="0"/>
27          <portSpacing port="sink_out 2" spacing="0"/>
28          <portSpacing port="sink_out 3" spacing="0"/>
29        </process>
30      </operator>
31      <portSpacing port="source_input 1" spacing="0"/>
32      <portSpacing port="sink_result 1" spacing="0"/>
33    </process>
34  </operator>
35 </process>
```

A.4 Wrapper

A.4.1 Naive Bayes

```
1
2  <?xml version="1.0" encoding="UTF-8" standalone="no"?>
3  <process version="5.3.013">
4    <context>
5      <input/>
6      <output/>
7      <macros/>
8    </context>
9    <operator activated="true" class="process" compatibility="5.3.013" expanded="true
       " name="Process">
10     <parameter key="logverbosity" value="all"/>
11     <parameter key="logfile" value="D:\testexrff.xrff"/>
12     <process expanded="true">
13       <operator activated="true" class="x_validation" compatibility="5.3.013"
           expanded="true" height="112" name="10 fold cross validation" width="90" x
           ="380" y="120">
14         <parameter key="use_local_random_seed" value="true"/>
15         <process expanded="true">
16           <operator activated="true" class="naive_bayes" compatibility="5.3.013"
               expanded="true" height="76" name="Naive Bayes" width="90" x="45" y="
               30"/>
17           <connect from_port="training" to_op="Naive Bayes" to_port="training set"
               />
18           <connect from_op="Naive Bayes" from_port="model" to_port="model"/>
19           <portSpacing port="source_training" spacing="0"/>
20           <portSpacing port="sink_model" spacing="0"/>
21           <portSpacing port="sink_through 1" spacing="0"/>
22         </process>
23         <process expanded="true">
24           <operator activated="true" class="apply_model" compatibility="5.3.013"
               expanded="true" height="76" name="Apply Model (11)" width="90" x="45"
                y="30">
25             <list key="application_parameters"/>
26           </operator>
27           <operator activated="true" class="performance" compatibility="5.3.013"
               expanded="true" height="76" name="Performance (11)" width="90" x="212
               " y="30"/>
28           <connect from_port="model" to_op="Apply Model (11)" to_port="model"/>
29           <connect from_port="test set" to_op="Apply Model (11)" to_port="
               unlabelled data"/>
30           <connect from_op="Apply Model (11)" from_port="labelled data" to_op="
               Performance (11)" to_port="labelled data"/>
```

```
31          <connect from_op="Performance (11)" from_port="performance" to_port="
                averagable 1"/>
32          <portSpacing port="source_model" spacing="0"/>
33          <portSpacing port="source_test set" spacing="0"/>
34          <portSpacing port="source_through 1" spacing="0"/>
35          <portSpacing port="sink_averagable 1" spacing="0"/>
36          <portSpacing port="sink_averagable 2" spacing="0"/>
37        </process>
38      </operator>
39      <portSpacing port="source_input 1" spacing="0"/>
40      <portSpacing port="sink_result 1" spacing="0"/>
41    </process>
42  </operator>
43 </process>
44  port="source_input 1" spacing="0"/>
45      <portSpacing port="sink_result 1" spacing="0"/>
46    </process>
47  </operator>
48 </process>
```

A.4.2 Random Forest

```
1
2  <?xml version="1.0" encoding="UTF-8" standalone="no"?>
3  <process version="5.3.013">
4    <context>
5      <input/>
6      <output/>
7      <macros/>
8    </context>
9    <operator activated="true" class="process" compatibility="5.3.013" expanded="true
         " name="Process">
10     <parameter key="logverbosity" value="all"/>
11     <parameter key="logfile" value="D:\testexrff.xrff"/>
12     <process expanded="true">
13       <operator activated="true" class="x_validation" compatibility="5.3.013"
             expanded="true" height="112" name="10 fold cross validation" width="90" x
             ="380" y="120">
14         <parameter key="use_local_random_seed" value="true"/>
15         <process expanded="true">
16           <operator activated="true" class="weka:W-RandomForest" compatibility="
                 5.3.000" expanded="true" height="76" name="W-RandomForest (7)" width=
                 "90" x="179" y="30">
17             <parameter key="I" value="500.0"/>
18             <parameter key="K" value="1.0"/>
19             <parameter key="depth" value="0"/>
```

```
20          </operator>
21          <connect from_port="training" to_op="W-RandomForest (7)" to_port="
                training set"/>
22          <connect from_op="W-RandomForest (7)" from_port="model" to_port="model"/>
23          <portSpacing port="source_training" spacing="0"/>
24          <portSpacing port="sink_model" spacing="0"/>
25          <portSpacing port="sink_through 1" spacing="0"/>
26        </process>
27        <process expanded="true">
28          <operator activated="true" class="apply_model" compatibility="5.3.013"
                expanded="true" height="76" name="Apply Model (11)" width="90" x="45"
                 y="30">
29            <list key="application_parameters"/>
30          </operator>
31          <operator activated="true" class="performance" compatibility="5.3.013"
                expanded="true" height="76" name="Performance (11)" width="90" x="212
                " y="30"/>
32          <connect from_port="model" to_op="Apply Model (11)" to_port="model"/>
33          <connect from_port="test set" to_op="Apply Model (11)" to_port="
                unlabelled data"/>
34          <connect from_op="Apply Model (11)" from_port="labelled data" to_op="
                Performance (11)" to_port="labelled data"/>
35          <connect from_op="Performance (11)" from_port="performance" to_port="
                averagable 1"/>
36          <portSpacing port="source_model" spacing="0"/>
37          <portSpacing port="source_test set" spacing="0"/>
38          <portSpacing port="source_through 1" spacing="0"/>
39          <portSpacing port="sink_averagable 1" spacing="0"/>
40          <portSpacing port="sink_averagable 2" spacing="0"/>
41        </process>
42      </operator>
43      <portSpacing port="source_input 1" spacing="0"/>
44      <portSpacing port="sink_result 1" spacing="0"/>
45    </process>
46  </operator>
47 </process>
```

A.5 Naive Bayes

```
1
2 <?xml version="1.0" encoding="UTF-8" standalone="no"?>
3 <process version="5.3.013">
4   <context>
5     <input/>
6     <output/>
```

```
7     <macros/>
8   </context>
9   <operator activated="true" class="process" compatibility="5.3.013" expanded="true
        " name="Process">
10    <parameter key="logverbosity" value="all"/>
11    <parameter key="logfile" value="D:\testexrff.xrff"/>
12    <process expanded="true">
13      <operator activated="true" class="x_validation" compatibility="5.3.013"
            expanded="true" height="112" name="10 fold validation" width="90" x="313"
             y="210">
14        <parameter key="use_local_random_seed" value="true"/>
15        <process expanded="true">
16          <operator activated="true" class="naive_bayes" compatibility="5.3.013"
                expanded="true" height="76" name="Naive Bayes (2)" width="90" x="112"
                 y="30"/>
17          <connect from_port="training" to_op="Naive Bayes (2)" to_port="training
                set"/>
18          <connect from_op="Naive Bayes (2)" from_port="model" to_port="model"/>
19          <portSpacing port="source_training" spacing="0"/>
20          <portSpacing port="sink_model" spacing="0"/>
21          <portSpacing port="sink_through 1" spacing="0"/>
22        </process>
23        <process expanded="true">
24          <operator activated="true" class="apply_model" compatibility="5.3.013"
                expanded="true" height="76" name="Apply Model (11)" width="90" x="112
                " y="30">
25            <list key="application_parameters"/>
26          </operator>
27          <operator activated="true" class="performance" compatibility="5.3.013"
                expanded="true" height="76" name="Performance (11)" width="90" x="246
                " y="30"/>
28          <connect from_port="model" to_op="Apply Model (11)" to_port="model"/>
29          <connect from_port="test set" to_op="Apply Model (11)" to_port="
                unlabelled data"/>
30          <connect from_op="Apply Model (11)" from_port="labelled data" to_op="
                Performance (11)" to_port="labelled data"/>
31          <connect from_op="Performance (11)" from_port="performance" to_port="
                averagable 1"/>
32          <portSpacing port="source_model" spacing="0"/>
33          <portSpacing port="source_test set" spacing="0"/>
34          <portSpacing port="source_through 1" spacing="0"/>
35          <portSpacing port="sink_averagable 1" spacing="0"/>
36          <portSpacing port="sink_averagable 2" spacing="0"/>
37        </process>
38      </operator>
39      <portSpacing port="source_input 1" spacing="0"/>
40      <portSpacing port="sink_result 1" spacing="0"/>
41    </process>
42  </operator>
```

```
</process>
```

A.6 Random Forest

```

<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<process version="5.3.013">
  <context>
    <input/>
    <output/>
    <macros/>
  </context>
  <operator activated="true" class="process" compatibility="5.3.013" expanded="true
      " name="Process">
    <parameter key="logverbosity" value="all"/>
    <parameter key="logfile" value="D:\testexrff.xrff"/>
    <process expanded="true">
      <operator activated="true" class="x_validation" compatibility="5.3.013"
          expanded="true" height="112" name="10 fold cross validation" width="90" x
          ="313" y="255">
        <parameter key="use_local_random_seed" value="true"/>
        <process expanded="true">
          <operator activated="true" class="weka:W-RandomForest" compatibility="
              5.3.000" expanded="true" height="76" name="W-RandomForest (7)" width=
              "90" x="133" y="30">
            <parameter key="I" value="1000.0"/>
            <parameter key="K" value="1.0"/>
            <parameter key="depth" value="0"/>
          </operator>
          <connect from_port="training" to_op="W-RandomForest (7)" to_port="
              training set"/>
          <connect from_op="W-RandomForest (7)" from_port="model" to_port="model"/>
          <portSpacing port="source_training" spacing="0"/>
          <portSpacing port="sink_model" spacing="0"/>
          <portSpacing port="sink_through 1" spacing="0"/>
        </process>
        <process expanded="true">
          <operator activated="true" class="apply_model" compatibility="5.3.013"
              expanded="true" height="76" name="Apply Model (11)" width="90" x="45"
               y="120">
            <list key="application_parameters"/>
          </operator>
          <operator activated="true" class="performance" compatibility="5.3.013"
              expanded="true" height="76" name="Performance (11)" width="90" x="214
              " y="75"/>
```

```
32          <connect from_port="model" to_op="Apply Model (11)" to_port="model"/>
33          <connect from_port="test set" to_op="Apply Model (11)" to_port="
                unlabelled data"/>
34          <connect from_op="Apply Model (11)" from_port="labelled data" to_op="
                Performance (11)" to_port="labelled data"/>
35          <connect from_op="Performance (11)" from_port="performance" to_port="
                averagable 1"/>
36          <portSpacing port="source_model" spacing="0"/>
37          <portSpacing port="source_test set" spacing="0"/>
38          <portSpacing port="source_through 1" spacing="0"/>
39          <portSpacing port="sink_averagable 1" spacing="0"/>
40          <portSpacing port="sink_averagable 2" spacing="0"/>
41        </process>
42      </operator>
43      <portSpacing port="source_input 1" spacing="0"/>
44      <portSpacing port="sink_result 1" spacing="0"/>
45    </process>
46  </operator>
47 </process>
```

A.7 SVM

Anexo B

Código de ILP

```
:-['episodio.pl'].
:-['intervencao.pl'].
:-['estatistica.pl'].
:-['cirurgia.pl'].
:-['consumo.pl'].
:- set(language, 1).
:- set(minpos, 10).
:- set(samplesize, 3).
:- set(nodes, 100000).
:- set(noise, 10).
:- set(record,true).



altura(_).

:-determination(internamento/1, cirurgia/9).
:-determination(internamento/1, intervencao/6).
:-determination(internamento/1, consumo/13).
:-determination(internamento/1, episodio/32).
:-determination(internamento/1, estatistica/21).

:- determination(internamento/1, lteq/2).
:- determination(internamento/1, gteq/2).
:- determination(internamento/1,foradistrito/1).
:- determination(internamento/1,nodia/1).
:- determination(internamento/1,gteqfreq/2).
:- determination(internamento/1,lteqfreq/2).
%:- determination(internamento/1,principiodosegteq/3).
%:- determination(internamento/1,principiodosegteq/3).
%:- determination(internamento/1,nomeciedosegteq/3).
%:- determination(internamento/1,nomeciedoselteq/3).
```

```
%:- determination(internamento/1,principioqtadmgteq/3).
%:- determination(internamento/1,principioqtadmlteq/3).
%:- determination(internamento/1,nomecieqtadmgteq/3).
%:- determination(internamento/1,nomeciedoselteq/3).
%:- determination(internamento/1,principioqtdiagteq/3).
%:- determination(internamento/1,principiodoselteq/3).
%:- determination(internamento/1,nomecieqtdiagteq/3).
%:- determination(internamento/1,nomecieqtadmlteq/3).

:- determination(internamento/1,conlteq/2).
:- determination(internamento/1,congteq/2).
:- determination(internamento/1,cirlteq/2).
:- determination(internamento/1,cirgteq/2).
:- determination(internamento/1,intlteq/2).
:- determination(internamento/1,intgteq/2).


:- modeb(1,conlteq(+episodioid,#real)).
:- modeb(1,congteq(+episodioid,#real)).

:- modeb(1,cirlteq(+episodioid,#real)).
:- modeb(1,cirgteq(+episodioid,#real)).

:- modeb(1,intlteq(+n_reg_oper,#real)).
:- modeb(1,intgteq(+n_reg_oper,#real)).

:- modeb(1,gteqfreq(+principio_activo,#freq)).
:- modeb(1,lteqfreq(+principio_activo,#freq)).

:- modeb(1,principiodosegteq(+principio_activo,-dose,#real)).
:- modeb(1,principiodoselteq(+principio_activo,-dose,#real)).


:- modeb(1,principiodosegteq(+principio_activo_freq,-media_dose_principio_activo,#
    real)).
:- modeb(1,principiodoselteq(+principio_activo_freq,-media_dose_principio_activo,#
    real)).


:- modeb(1,nomeciedosegteq(+nome_cient,-dose,#real)).
:- modeb(1,nomeciedoselteq(+nome_cient,-dose,#real)).

:- modeb(1,nomeciedosegteq(+nome_cient_freq,-media_dose_nome_cient,#real)).
:- modeb(1,nomeciedoselteq(+nome_cient_freq,-media_dose_nome_cient,#real)).

:- modeb(1,principioqtadmgteq(+principio_activo,-qt_adm,#real)).
:- modeb(1,principioqtadmlteq(+principio_activo,-qt_adm,#real)).

```

```
79  :- modeb(1,principioqtadmgteq(+principio_activo_freq,-media_qt_adm_principio_activo
        ,#real)).
80  :- modeb(1,principioqtadmlteq(+principio_activo_freq,-media_qt_adm_principio_activo
        ,#real)).
81
82  :- modeb(1,nomecieqtadmgteq(+nome_cient,-qt_adm,#real)).
83  :- modeb(1,nomecieqtadmlteq(+nome_cient,-qt_adm,#real)).
84
85  :- modeb(1,nomecieqtadmgteq(+nome_cient_freq,-media_qt_adm_principio_activo,#real))
        .
86  :- modeb(1,nomecieqtadmlteq(+nome_cient_freq,-media_qt_adm_principio_activo,#real))
        .
87
88
89  :- modeb(1,principioqtdiagteq(+principio_activo,-qt_dia,#real)).
90  :- modeb(1,principioqtdialteq(+principio_activo,-qt_dia,#real)).
91
92  :- modeb(1,principioqtdiagteq(+principio_activo_freq,-media_qt_dia_principio_activo
        ,#real)).
93  :- modeb(1,principioqtdialteq(+principio_activo_freq,-media_qt_dia_principio_activo
        ,#real)).
94
95
96  :- modeb(1,nomecieqtdiagteq(+nome_cient,-qt_dia,#real)).
97  :- modeb(1,nomecieqtdialteq(+nome_cient,-qt_dia,#real)).
98
99  :- modeb(1,nomecieqtdiagteq(+nome_cient_freq,-media_qt_dia_principio_activo,#real))
        .
100 :- modeb(1,nomecieqtdialteq(+nome_cient_freq,-media_qt_dia_principio_activo,#real))
        .
101
102
103
104
105 :- modeb(1,gteq(+tempointer,#real)).
106 :- modeb(1,lteq(+tempointer,#real)).
107 :- modeb(1,gteq(+idade,#real)).
108 :- modeb(1,lteq(+idade,#real)).
109 :- modeb(1,gteq(+qt_adm,#real)).
110 :- modeb(1,lteq(+qt_adm,#real)).
111 :- modeb(1,gteq(+qt_dia,#real)).
112 :- modeb(1,lteq(+qt_dia,#real)).
113 :- modeb(1,lteq(+qt_dia,#real)).
114
115 :- modeb(1,gteq(+principio_activo,-dose,#real)).
116 :- modeb(1,lteq(+principio_activo,-dose,#real)).
117
118 :- modeb(1,gteq(+altura,#real)).
119 :- modeb(1,lteq(+altura,#real)).
```

```
120
121
122 :- modeb(1,gteq(+interv_dt_oper_sec1,#real)).
123 :- modeb(1,lteq(+interv_dt_oper_sec1,#real)).
124
125 :- modeb(1,gteq(+interv_dt_oper_sec2,#real)).
126 :- modeb(1,lteq(+interv_dt_oper_sec2,#real)).
127
128 :- modeb(1,gteq(+media_qt_adm_principio_activo,#real)).
129 :- modeb(1,lteq(+media_qt_adm_principio_activo,#real)).
130
131 :- modeb(1,gteq(+media_qt_dia_principio_activo,#real)).
132 :- modeb(1,lteq(+media_qt_dia_principio_activo,#real)).
133
134 :- modeb(1,gteq(+media_dose_nome_cient,#real)).
135 :- modeb(1,lteq(+media_dose_nome_cient,#real)).
136
137 :- modeb(1,gteq(+media_qt_dia_nome_cient,#real)).
138 :- modeb(1,lteq(+media_qt_dia_nome_cient,#real)).
139
140 :- modeb(1,gteq(+media_qt_adm_nome_cient,#real)).
141 :- modeb(1,lteq(+media_qt_adm_nome_cient,#real)).
142
143 :- modeb(1,gteq(+media_qtd_duracao_anestesia,#real)).
144 :- modeb(1,lteq(+media_qtd_duracao_anestesia,#real)).
145
146 :- modeb(1,gteq(+media_qtd_duracao_cirurgia,#real)).
147 :- modeb(1,lteq(+media_qtd_duracao_cirurgia,#real)).
148
149
150 :- modeb(1,gteq(+media_qtd_duracao_cirurgia,#real)).
151 :- modeb(1,lteq(+media_qtd_duracao_cirurgia,#real)).
152
153
154 :- modeb(1,gteq(+qtd_duracao_anestesia,#real)).
155 :- modeb(1,lteq(+qtd_duracao_anestesia,#real)).
156
157 :- modeb(1,gteq(+qtd_duracao_cirurgia,#real)).
158 :- modeb(1,lteq(+qtd_duracao_cirurgia,#real)).
159
160 :- modeb(1,gteq(+n_ord,#real)).
161 :- modeb(1,lteq(+n_ord,#real)).
162
163
164 :- modeb(1,gteq(+dose,#real)).
165 :- modeb(1,lteq(+dose,#real)).
166
167 :- modeb(1,foradistrito(+episodioid)).
168 :- modeb(1,nodia(+episodioid)).
```

```
169
170
171 :-modeh(1,internamento(+episodioid)).
172 :-modeb(1,episodio(
173      -altura,
174      -tempointer,
175      -t_doente,
176      -t_episodio,
177       +episodioid
178       ,-dt_int,
179       -dt_alta,
180       -cod_doente
181       ,-centro_saude
182       ,-pais
183       ,-distrito
184       ,-concelho
185       ,-freguesia
186       ,-localidade
187       ,-pais_nasc
188       ,-distrito_nasc
189       ,-concelho_nasc
190       ,-freguesia_nasc
191       ,-nacionalidade
192       ,-naturalidade
193       ,-escolaridade
194       ,-estado_civil
195       ,-estado_profissao
196       ,-profissao
197       ,-flag_prog_ctrl_hiv
198       ,-grupo_sang
199       ,-medico_familia
200       ,-sexo
201       ,-dt_nasc
202       ,-doente
203       ,-cod_aplicacao
204      ,-idade)).
205
206
207 :-modeb(*,episodio(
208      -altura,
209      -tempointer,
210      -t_doente,
211      -t_episodio,
212       +episodioid
213       ,-dt_int,
214       -dt_alta,
215       -cod_doente
216       ,-centro_saude
217       ,-pais
```

```
      ,-distrito
      ,-concelho
      ,-freguesia
      ,-localidade
      ,-pais_nasc
      ,-distrito_nasc
      ,-concelho_nasc
      ,-freguesia_nasc
      ,#nacionalidade
      ,-naturalidade
      ,-escolaridade
      ,-estado_civil
      ,-estado_profissao
      ,-profissao
      ,-flag_prog_ctrl_hiv
      ,-grupo_sang
      ,-medico_familia
      ,-sexo
      ,-dt_nasc
      ,-doente
      ,-cod_aplicacao
     ,-idade)).


:-modeb(*,episodio(
     -altura,
     -tempointer,
     -t_doente,
     -t_episodio,
      +episodioid
      ,-dt_int,
      -dt_alta,
      -cod_doente
      ,-centro_saude
      ,-pais
      ,-distrito
      ,-concelho
      ,-freguesia
      ,-localidade
      ,-pais_nasc
      ,-distrito_nasc
      ,-concelho_nasc
      ,-freguesia_nasc
      ,-nacionalidade
      ,-naturalidade
      ,-escolaridade
      ,-estado_civil
      ,-estado_profissao
```

```
267       ,-profissao
268       ,-flag_prog_ctrl_hiv
269       ,-grupo_sang
270       ,-medico_familia
271       ,#sexo
272       ,-dt_nasc
273       ,-doente
274       ,-cod_aplicacao
275      ,-idade)).
276
277 :-modeb(*,episodio(
278      -altura,
279      -tempointer,
280      -t_doente,
281      -t_episodio,
282       +episodioid
283       ,-dt_int,
284       -dt_alta,
285       -cod_doente
286       ,-centro_saude
287       ,-pais
288       ,-distrito
289       ,-concelho
290       ,-freguesia
291       ,-localidade
292       ,#pais_nasc
293       ,-distrito_nasc
294       ,-concelho_nasc
295       ,-freguesia_nasc
296       ,-nacionalidade
297       ,-naturalidade
298       ,-escolaridade
299       ,-estado_civil
300       ,-estado_profissao
301       ,-profissao
302       ,-flag_prog_ctrl_hiv
303       ,-grupo_sang
304       ,-medico_familia
305       ,-sexo
306       ,-dt_nasc
307       ,-doente
308       ,-cod_aplicacao
309      ,-idade)).
310
311
312
313 :-modeb(*,cirurgia(
314       +episodioid
315       ,-n_reg_oper
```

```
      ,-descricao_anestesia
      ,-cod_asa
      ,-tipo_anestesia
      ,-tipo_cir
      ,-tipo_assepsia
      ,-qtd_duracao_anestesia
      ,-qtd_duracao_cirurgia
 )).



:-modeb(*,cirurgia(
      +episodioid
      ,-n_reg_oper
      ,-descricao_anestesia
      ,-cod_asa
      ,-tipo_anestesia
      ,#tipo_cir
      ,-tipo_assepsia
      ,-qtd_duracao_anestesia
      ,-qtd_duracao_cirurgia
 )).

:-modeb(*,cirurgia(
      +episodioid
      ,-n_reg_oper
      ,#descricao_anestesia
      ,-cod_asa
      ,-tipo_anestesia
      ,-tipo_cir
      ,-tipo_assepsia
      ,-qtd_duracao_anestesia
      ,-qtd_duracao_cirurgia
 )).

 :-modeb(*,cirurgia(
      +episodioid
      ,-n_reg_oper
      ,-descricao_anestesia
      ,-cod_asa
      ,-tipo_anestesia
      ,-tipo_cir
      ,#tipo_assepsia
      ,-qtd_duracao_anestesia
      ,-qtd_duracao_cirurgia
 )).
:-modeb(*,cirurgia(
```

```
365       +episodioid
366       ,-n_reg_oper
367       ,-descricao_anestesia
368       ,-cod_asa
369       ,-tipo_anestesia
370       ,-tipo_cir
371       ,#tipo_assepsia
372       ,-qtd_duracao_anestesia
373       ,-qtd_duracao_cirurgia
374  )).
375
376 :-modeb(20,consumo(
377       +episodioid
378       ,-qt_adm
379       ,-qt_dia
380       ,-administracao
381       ,-nome_cient
382       ,-nome_comercial
383       ,-principio_activo
384       ,-dose
385       ,-via_adm
386       ,-freq
387       ,-unid_med
388       ,-horario
389       ,-prescrito)).
390
391
392 :-modeb(20,consumo(
393       +episodioid
394       ,-qt_adm
395       ,-qt_dia
396       ,-administracao
397       ,-nome_cient
398       ,-nome_comercial
399       ,#principio_activo
400       ,-dose
401       ,-via_adm
402       ,-freq
403       ,-unid_med
404       ,-horario
405       ,-prescrito)).
406
407  :-modeb(20,consumo(
408       +episodioid
409       ,-qt_adm
410       ,-qt_dia
411       ,-administracao
412       ,-nome_cient
413       ,-nome_comercial
```

```
414       ,-principio_activo
415       ,-dose
416       ,-via_adm
417       ,-freq
418       ,-unid_med
419       ,-horario
420       ,#prescrito)).
421
422 :-modeb(20,consumo(
423       +episodioid
424       ,-qt_adm
425       ,-qt_dia
426       ,-administracao
427       ,-nome_cient
428       ,-nome_comercial
429       ,-principio_activo
430       ,-dose
431       ,-via_adm
432       ,#freq
433       ,-unid_med
434       ,-horario
435       ,-prescrito)).
436
437 :-modeb(*,intervencao(
438       +n_reg_oper
439       ,-descricao
440       ,-codificacao
441       ,-flag_principal
442       ,-n_ord
443       ,-dt_oper)).
444
445
446 :-modeb(*,intervencao(
447       +n_reg_oper
448       ,#descricao
449       ,-codificacao
450       ,-flag_principal
451       ,-n_ord
452       ,-dt_oper)).
453
454 :-modeb(*,intervencao(
455       +n_reg_oper
456       ,-descricao
457       ,#codificacao
458       ,-flag_principal
459       ,-n_ord
460       ,-dt_oper)).
461
462   :-modeb(*,estatistica(
```

```
463     -interv_dt_oper_principal,
464     -interv_dt_oper_sec1,
465     -interv_dt_oper_sec2,
466     -media_dose_principio_activo,
467     -media_qt_adm_principio_activo,
468     -media_qt_dia_principio_activo,
469     -media_dose_nome_cient,
470    - media_qt_dia_nome_cient,
471    - media_qt_adm_nome_cient,
472     -media_qtd_duracao_anestesia,
473     -media_qtd_duracao_cirurgia,
474     +episodioid,
475     -nome_cient_freq,
476     -principio_activo_freq,
477     -nome_comercial_freq,
478     -via_adm_freq,
479    - horario_freq,
480    -adm_freq,
481    - interv_desc_principal,
482     -interv_desc_sec1,
483    - interv_desc_sec2)).
484
485
486
487 :-modeb(*,estatistica(
488     -interv_dt_oper_principal,
489     -interv_dt_oper_sec1,
490     -interv_dt_oper_sec2,
491     -media_dose_principio_activo,
492     -media_qt_adm_principio_activo,
493     -media_qt_dia_principio_activo,
494     -media_dose_nome_cient,
495    - media_qt_dia_nome_cient,
496    - media_qt_adm_nome_cient,
497     -media_qtd_duracao_anestesia,
498     -media_qtd_duracao_cirurgia,
499     +episodioid,
500     -nome_cient_freq,
501     -principio_activo_freq,
502     -nome_comercial_freq,
503     -via_adm_freq,
504    - horario_freq,
505    -adm_freq,
506     #interv_desc_principal,
507     -interv_desc_sec1,
508    - interv_desc_sec2)).
509
510
511 :-modeb(*,estatistica(
```

```
512     -interv_dt_oper_principal,
513     -interv_dt_oper_sec1,
514     -interv_dt_oper_sec2,
515     #media_dose_principio_activo,
516     -media_qt_adm_principio_activo,
517     -media_qt_dia_principio_activo,
518     -media_dose_nome_cient,
519     -media_qt_dia_nome_cient,
520     -media_qt_adm_nome_cient,
521     -media_qtd_duracao_anestesia,
522     -media_qtd_duracao_cirurgia,
523     +episodioid,
524     -nome_cient_freq,
525     #principio_activo_freq,
526     -nome_comercial_freq,
527     -via_adm_freq,
528     -horario_freq,
529     -adm_freq,
530     -interv_desc_principal,
531     -interv_desc_sec1,
532     -interv_desc_sec2)).
533
534
535 :-modeb(*,estatistica(
536     -interv_dt_oper_principal,
537     -interv_dt_oper_sec1,
538     -interv_dt_oper_sec2,
539     -media_dose_principio_activo,
540     -media_qt_adm_principio_activo,
541     -media_qt_dia_principio_activo,
542     -media_dose_nome_cient,
543     -media_qt_dia_nome_cient,
544     -media_qt_adm_nome_cient,
545     -media_qtd_duracao_anestesia,
546     -media_qtd_duracao_cirurgia,
547     +episodioid,
548     -nome_cient_freq,
549     #principio_activo_freq,
550     -nome_comercial_freq,
551     -via_adm_freq,
552     -horario_freq,
553     -adm_freq,
554     -interv_desc_principal,
555     -interv_desc_sec1,
556     -interv_desc_sec2)).
557
558 -modeb(*,estatistica(
559     -interv_dt_oper_principal,
560     -interv_dt_oper_sec1,
```

```
    -interv_dt_oper_sec2,
    -media_dose_principio_activo,
    -media_qt_adm_principio_activo,
    -media_qt_dia_principio_activo,
    #media_dose_nome_cient,
   - media_qt_dia_nome_cient,
   - media_qt_adm_nome_cient,
    -media_qtd_duracao_anestesia,
    -media_qtd_duracao_cirurgia,
    +episodioid,
    #nome_cient_freq,
    -principio_activo_freq,
    -nome_comercial_freq,
    -via_adm_freq,
   - horario_freq,
   -adm_freq,
   - interv_desc_principal,
    -interv_desc_sec1,
   - interv_desc_sec2)).



:-modeb(*,estatistica(
    -interv_dt_oper_principal,
    -interv_dt_oper_sec1,
    -interv_dt_oper_sec2,
    -media_dose_principio_activo,
    -media_qt_adm_principio_activo,
    -media_qt_dia_principio_activo,
    -media_dose_nome_cient,
    -media_qt_dia_nome_cient,
    -media_qt_adm_nome_cient,
    -media_qtd_duracao_anestesia,
    -media_qtd_duracao_cirurgia,
    +episodioid,
    #nome_cient_freq,
    -principio_activo_freq,
    -nome_comercial_freq,
    -via_adm_freq,
    -horario_freq,
    -adm_freq,
    -interv_desc_principal,
    -interv_desc_sec1,
    -interv_desc_sec2)).

% background knowledge

gteq(X,Y):-
```

```
610   number(X), number(Y),
611   X >= Y, !.
612 gteq(X,X):-
613   number(X).
614
615 lteq(X,Y):-
616   number(X), number(Y),
617   X =< Y, !.
618 lteq(X,X):-
619   number(X).
620
621
622
623
624 foradistrito(A):- episodio(_,_,_,_,A,_,_,_,_,_,_,B,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_),
625 B \='braga',
626 B \='desconhecido',!.
627
628
629 nodia(A):- episodio(_,_,_,_,A,B,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_,_),cirurgia(A,C,_,_,_,_,_,_,_),intervencao(C,_,_,_,_,E),
630 B == E,
631 B \='desconhecido',!.
632
633
634
635 %principios activos
636 principiodosegteq(F,X,Y):-
637 consumo(_,_,_,_,_,_,F,X,_,_,_,_,_,_),number(X), number(Y),
638   X >= Y, !.
639
640 principiodoselteq(F,X,Y):-
641 consumo(_,_,_,_,_,_,F,X,_,_,_,_,_,_),number(X), number(Y),
642   X  =< Y, !.
643
644
645 principiodosegteq(F,X,X):-
646 consumo(_,_,_,_,_,_,F,X,_,_,_,_,_,_),number(X).
647
648 principiodoselteq(F,X,X):-
649 consumo(_,_,_,_,_,_,F,X,_,_,_,_,_,_),number(X).
650
651
652
653
654
655
656 principioqtadmgteq(F,X,Y):-
```

```
consumo(_,X,_,_,_,_,F,_,_,_,_,_,_),number(X), number(Y),
  X >= Y, !.

principioqtadmlteq(F,X,Y):-
consumo(_,X,_,_,_,_,F,_,_,_,_,_,_),number(X), number(Y),
  X  =< Y, !.

principioqtadmgteq(F,X,X):-
consumo(_,X,_,_,_,_,F,_,_,_,_,_,_),number(X) .

principioqtadmlteq(F,X,X):-
consumo(_,X,_,_,_,_,F,_,_,_,_,_,_),number(X).




principioqtdiagteq(F,X,Y):-
consumo(_,_,X,_,_,_,F,_,_,_,_,_,_),number(X), number(Y),
  X >= Y, !.

principioqtdialteq(F,X,Y):-
consumo(_,_,X,_,_,_,F,_,_,_,_,_,_),number(X), number(Y),
  X  =< Y, !.

principioqtdiagteq(F,X,X):-
consumo(_,_,X,_,_,_,F,_,_,_,_,_,_),number(X).

principioqtdialteq(F,X,X):-
consumo(_,_,X,_,_,_,F,_,_,_,_,_,_),number(X).




%nome_cient

nomeciedosegteq(F,X,Y):-
consumo(_,_,_,_,_,F,_,X,_,_,_,_,_),number(X), number(Y),
  X >= Y, !.

nomeciedoselteq(F,X,Y):-
consumo(_,_,_,_,_,F,_,X,_,_,_,_,_),number(X), number(Y),
  X  =< Y, !.



nomeciedosegteq(F,X,X):-
consumo(_,_,_,_,_,F,_,X,_,_,_,_,_),number(X).

nomeciedoselteq(F,X,X):-
```

```
706 consumo(_,_,_,_,_,F,_,X,_,_,_,_,_,_),number(X).
707
708
709
710
711 nomecieqtadmgteq(F,X,Y):-
712 consumo(_,X,_,_,_,F,_,_,_,_,_,_,_,_),number(X), number(Y),
713   X >= Y, !.
714
715 nomecieqtadmlteq(F,X,Y):-
716 consumo(_,X,_,_,_,F,_,_,_,_,_,_,_,_),number(X), number(Y),
717   X  =< Y, !.
718
719 nomecieqtadmgteq(F,X,X):-
720 consumo(_,X,_,_,_,F,_,_,_,_,_,_,_,_),number(X).
721
722 nomecieqtadmlteq(F,X,X):-
723 consumo(_,X,_,_,_,F,_,_,_,_,_,_,_,_),number(X).
724
725 nomecieqtdiagteq(F,X,Y):-
726 consumo(_,_,X,_,_,F,_,_,_,_,_,_,_,_),number(X), number(Y),
727   X >= Y, !.
728
729 nomecieqtdialteq(F,X,Y):-
730 consumo(_,_,X,_,_,F,_,_,_,_,_,_,_,_),number(X), number(Y),
731   X  =< Y, !.
732
733 nomecieqtdiagteq(F,X,X):-
734 consumo(_,_,X,_,_,F,_,_,_,_,_,_,_,_),number(X).
735
736 nomecieqtdialteq(F,X,X):-
737 consumo(_,_,X,_,_,F,_,_,_,_,_,_,_,_),number(X).
738
739
740
741
742 conlteq(T,Y):-consumo(T,_,_,_,_,_,_,_,_,_,_,_,_,_),count(consumo(T,_,_,_,_,_,_,_,_,_,
    _,_,_),F), number(F),number(Y),F =< Y, !.
743 congteq(T,Y):-consumo(T,_,_,_,_,_,_,_,_,_,_,_,_,_),count(consumo(T,_,_,_,_,_,_,_,_,_,
    _,_,_),F),number(F), number(Y),F >= Y, !.
744
745
746 conlteq(T,F):-consumo(T,_,_,_,_,_,_,_,_,_,_,_,_,_),count(consumo(T,_,_,_,_,_,_,_,_,_,
    _,_,_),F),number(F).
747 congteq(T,F):-consumo(T,_,_,_,_,_,_,_,_,_,_,_,_,_),count(consumo(T,_,_,_,_,_,_,_,_,_,
    _,_,_),F),number(F) .
748
749
```

```
750 cirlteq(T,Y):-cirurgia(T,_,_,_,_,_,_,_,_),count(cirurgia(T,_,_,_,_,_,_,_,_),F),
        number(F),number(Y),F =< Y, !.
751 cirgteq(T,Y):-cirurgia(T,_,_,_,_,_,_,_,_),count(cirurgia(T,_,_,_,_,_,_,_,_),F),
        number(F),number(Y),F >= Y, !.
752
753 cirlteq(T,F):-cirurgia(T,_,_,_,_,_,_,_,_),count(cirurgia(T,_,_,_,_,_,_,_,_),F),
        number(F) .
754 cirgteq(T,F):-cirurgia(T,_,_,_,_,_,_,_,_),count(cirurgia(T,_,_,_,_,_,_,_,_),F),
        number(F) .
755
756
757 intlteq(T,Y):-intervencao(T,_,_,_,_,_),count(intervencao(T,_,_,_,_,_),F),number(F),
        number(Y) , F =< Y, !.
758 intgteq(T,Y):-intervencao(T,_,_,_,_,_),count(intervencao(T,_,_,_,_,_),F),number(F),
         number(Y),F >= Y, !.
759
760 intlteq(T,F):-intervencao(T,_,_,_,_,_),count(intervencao(T,_,_,_,_,_),F),number(F).
761 intgteq(T,F):-intervencao(T,_,_,_,_,_),count(intervencao(T,_,_,_,_,_),F),number(F).
762
763
764
765
766 count(P,Count) :-
767          findall(_,P,L),
768          length(L,Count).
769
770
771
772 gteqfreq(F,E):-
773 consumo(_,_,_,C,_,_,F,_,_,E,_,_,_),
774 consumo(_,_,_,D,_,_,F,_,_,G,_,_,_),C\=D,gteqf(E,G).
775
776 lteqfreq(F,E):-
777 consumo(_,_,_,C,_,_,F,_,_,E,_,_,_),
778 consumo(_,_,_,D,_,_,F,_,_,G,_,_,_),C\=D,lteqf(E,G).
779
780
781
782
783
784 freqtier1('toma unica').
785 freqtier2('1 x dia').
786 freqtier3('2 xs dia').
787 freqtier4('3 xs dia').
788 freqtier5('4 xs dia').
789 freqtier6('4/4 h').
790 freqtier7('2/2h').
791 freqtier8('perfusao').
792
```

```
793
794 gteqf(A,B):-freqtier1(A),freqtier1(B).
795 gteqf(A,B):-freqtier2(A),freqtier1(B).
796 gteqf(A,B):-freqtier3(A),freqtier1(B).
797 gteqf(A,B):-freqtier4(A),freqtier1(B).
798 gteqf(A,B):-freqtier5(A),freqtier1(B).
799 gteqf(A,B):-freqtier6(A),freqtier1(B).
800 gteqf(A,B):-freqtier7(A),freqtier1(B).
801 gteqf(A,B):-freqtier8(A),freqtier1(B).
802
803 gteqf(A,B):-freqtier2(A),freqtier2(B).
804 gteqf(A,B):-freqtier3(A),freqtier2(B).
805 gteqf(A,B):-freqtier4(A),freqtier2(B).
806 gteqf(A,B):-freqtier5(A),freqtier2(B).
807 gteqf(A,B):-freqtier6(A),freqtier2(B).
808 gteqf(A,B):-freqtier7(A),freqtier2(B).
809 gteqf(A,B):-freqtier8(A),freqtier2(B).
810
811 gteqf(A,B):-freqtier3(A),freqtier3(B).
812 gteqf(A,B):-freqtier4(A),freqtier3(B).
813 gteqf(A,B):-freqtier5(A),freqtier3(B).
814 gteqf(A,B):-freqtier6(A),freqtier3(B).
815 gteqf(A,B):-freqtier7(A),freqtier3(B).
816 gteqf(A,B):-freqtier8(A),freqtier3(B).
817
818 gteqf(A,B):-freqtier4(A),freqtier4(B).
819 gteqf(A,B):-freqtier5(A),freqtier4(B).
820 gteqf(A,B):-freqtier5(A),freqtier4(B).
821 gteqf(A,B):-freqtier6(A),freqtier4(B).
822 gteqf(A,B):-freqtier7(A),freqtier4(B).
823 gteqf(A,B):-freqtier8(A),freqtier4(B).
824
825 gteqf(A,B):-freqtier5(A),freqtier5(B).
826 gteqf(A,B):-freqtier6(A),freqtier5(B).
827 gteqf(A,B):-freqtier7(A),freqtier5(B).
828 gteqf(A,B):-freqtier8(A),freqtier5(B).
829
830 gteqf(A,B):-freqtier6(A),freqtier6(B).
831 gteqf(A,B):-freqtier7(A),freqtier6(B).
832 gteqf(A,B):-freqtier8(A),freqtier6(B).
833
834 gteqf(A,B):-freqtier7(A),freqtier7(B).
835 gteqf(A,B):-freqtier8(A),freqtier7(B).
836
837 gteqf(A,B):-freqtier8(A),freqtier8(B).
838
839
840
841
```

```
lteqf(B,A):-freqtier1(A),freqtier1(B).
lteqf(B,A):-freqtier2(A),freqtier1(B).
lteqf(B,A):-freqtier3(A),freqtier1(B).
lteqf(B,A):-freqtier4(A),freqtier1(B).
lteqf(B,A):-freqtier5(A),freqtier1(B).
lteqf(B,A):-freqtier6(A),freqtier1(B).
lteqf(B,A):-freqtier7(A),freqtier1(B).
lteqf(B,A):-freqtier8(A),freqtier1(B).


lteqf(B,A):-freqtier2(A),freqtier2(B).
lteqf(B,A):-freqtier3(A),freqtier2(B).
lteqf(B,A):-freqtier4(A),freqtier2(B).
lteqf(B,A):-freqtier5(A),freqtier2(B).
lteqf(B,A):-freqtier6(A),freqtier2(B).
lteqf(B,A):-freqtier7(A),freqtier2(B).
lteqf(B,A):-freqtier8(A),freqtier2(B).

lteqf(B,A):-freqtier3(A),freqtier3(B).
lteqf(B,A):-freqtier4(A),freqtier3(B).
lteqf(B,A):-freqtier5(A),freqtier3(B).
lteqf(B,A):-freqtier6(A),freqtier3(B).
lteqf(B,A):-freqtier7(A),freqtier3(B).
lteqf(B,A):-freqtier8(A),freqtier3(B).


lteqf(B,A):-freqtier4(A),freqtier4(B).
lteqf(B,A):-freqtier5(A),freqtier4(B).
lteqf(B,A):-freqtier5(A),freqtier4(B).
lteqf(B,A):-freqtier6(A),freqtier4(B).
lteqf(B,A):-freqtier7(A),freqtier4(B).
lteqf(B,A):-freqtier8(A),freqtier4(B).

lteqf(B,A):-freqtier5(A),freqtier5(B).
lteqf(B,A):-freqtier6(A),freqtier5(B).
lteqf(B,A):-freqtier7(A),freqtier5(B).
lteqf(B,A):-freqtier8(A),freqtier5(B).


lteqf(B,A):-freqtier6(A),freqtier6(B).
lteqf(B,A):-freqtier7(A),freqtier6(B).
lteqf(B,A):-freqtier8(A),freqtier6(B).

lteqf(B,A):-freqtier7(A),freqtier7(B).
lteqf(B,A):-freqtier8(A),freqtier7(B).

lteqf(B,A):-freqtier8(A),freqtier8(B).
```

Anexo C

Resultados

C.1 Naive Bayes

accuracy: 65.22% +/- 2.07% (mikro: 65.22%)			
	true S	true N	class precision
pred. S	1127	531	67.97%
pred. N	843	1449	63.22%
class recall	57.21%	73.18%	

Figura C.1: Matriz de confusão Naive Bayes

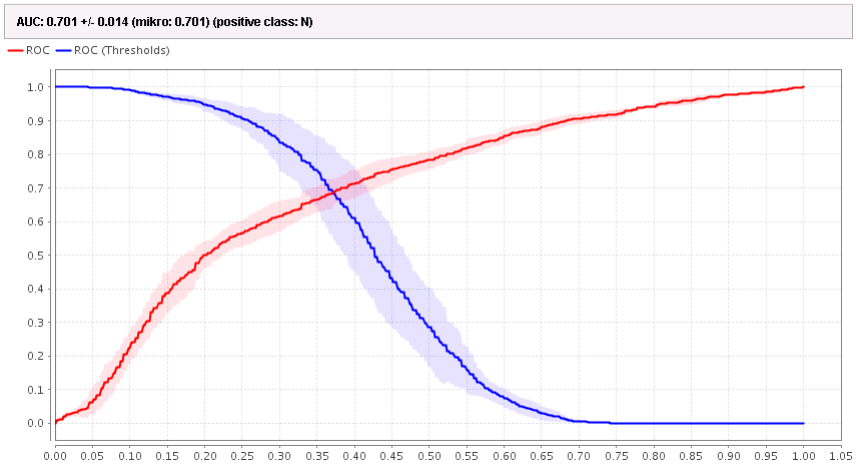

Figura C.2: ROC de Naive Bayes

C.2 Random Forest

C.2.1 50

accuracy: 67.52% +/- 1.71% (mikro: 67.52%)			
	true S	true N	class precision
pred. S	1371	684	66.72%
pred. N	599	1296	68.39%
class recall	69.59%	65.45%	

Figura C.3: Matriz de confusão de Random Forest 50 árvores

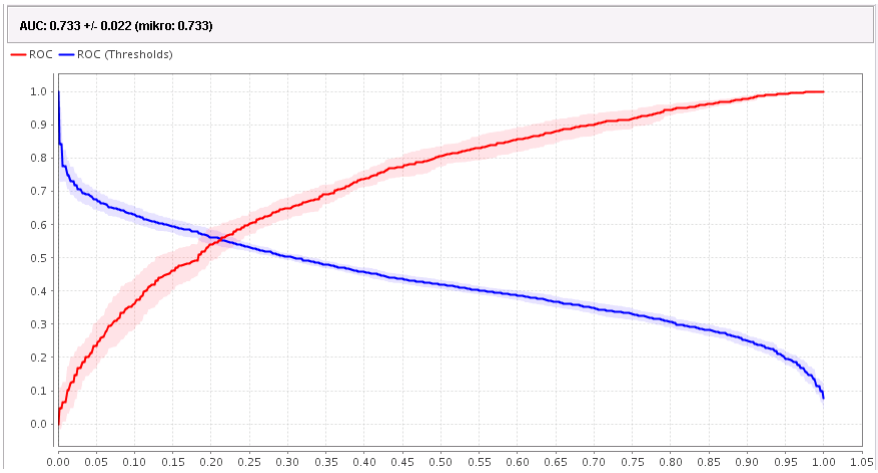

Figura C.4: ROC de Random Forest 50 árvores

C.2.2 100

accuracy: 67.72% +/- 2.84% (mikro: 67.72%)			
	true S	true N	class precision
pred. S	1384	689	66.76%
pred. N	586	1291	68.78%
class recall	70.25%	65.20%	

Figura C.5: Matriz de confusão de Random Forest de 100 árvores

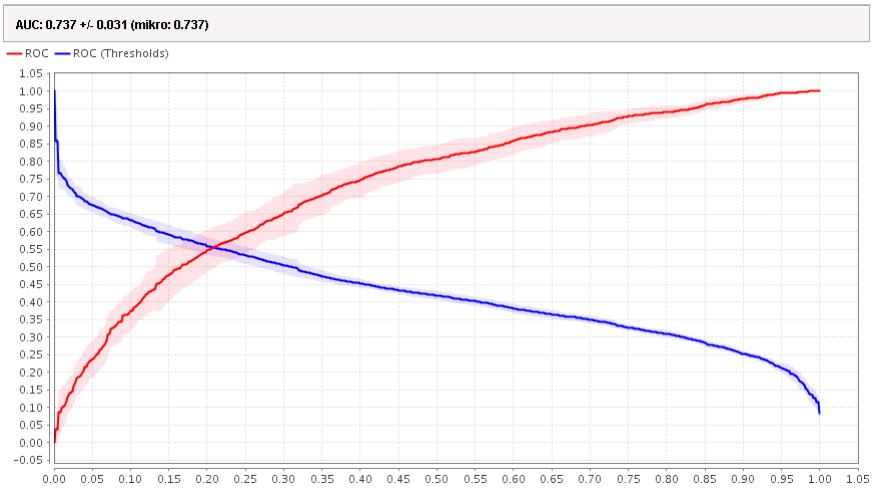

Figura C.6: ROC de Random Forest 100 árvores

C.2.3 250

accuracy: 67.75% +/- 1.02%

	true S	true N	class precision
pred. S	1378	682	66.89%
pred. N	592	1298	68.68%
class recall	69.95%	65.56%	

Figura C.7: Matriz de confusão de Random Forest de 250 árvores

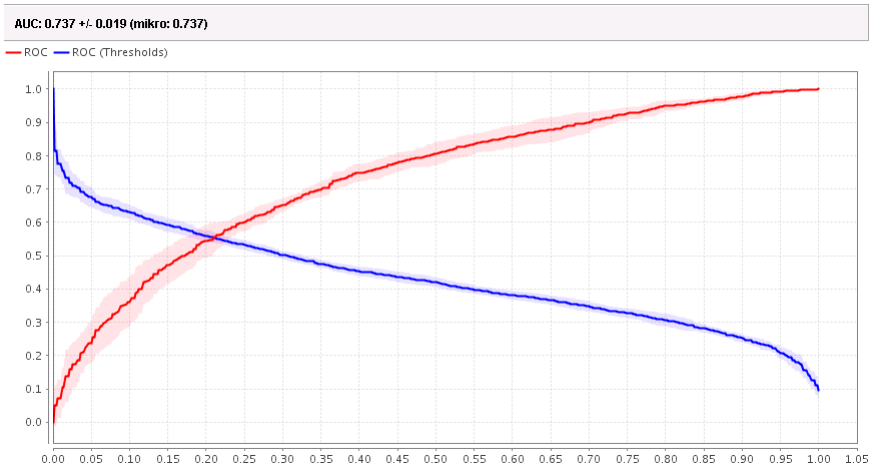

Figura C.8: ROC de Random Forest 250 árvores

C.2.4 500

accuracy: 67.97% +/- 0.98%			
	true S	true N	class precision
pred. S	1389	684	67.00%
pred. N	581	1296	69.05%
class recall	70.51%	65.45%	

Figura C.9: Matriz de confusão de Random Forest de 500 árvores

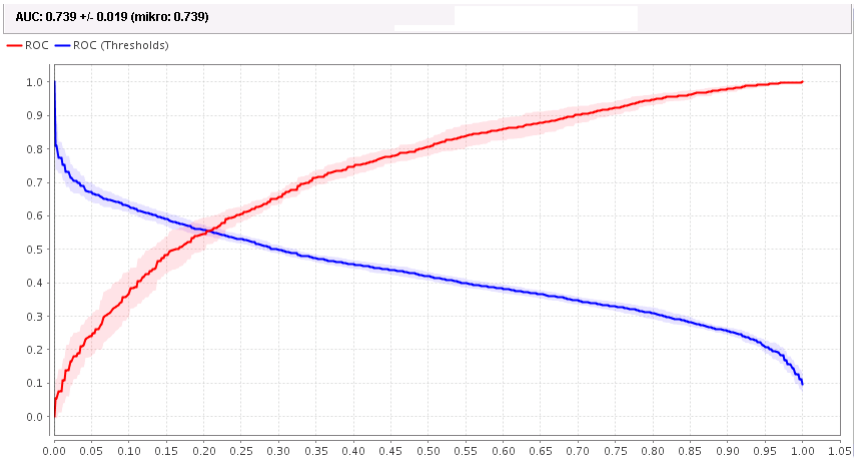

Figura C.10: ROC de Random Forest 500 árvores

C.2.5 1000

accuracy: 68.00% +/- 0.96%			
	true S	true N	class precision
pred. S	1395	689	66.94%
pred. N	575	1291	69.19%
class recall	70.81%	65.20%	

Figura C.11: Matriz de confusão de Random Forest de 1000 árvores

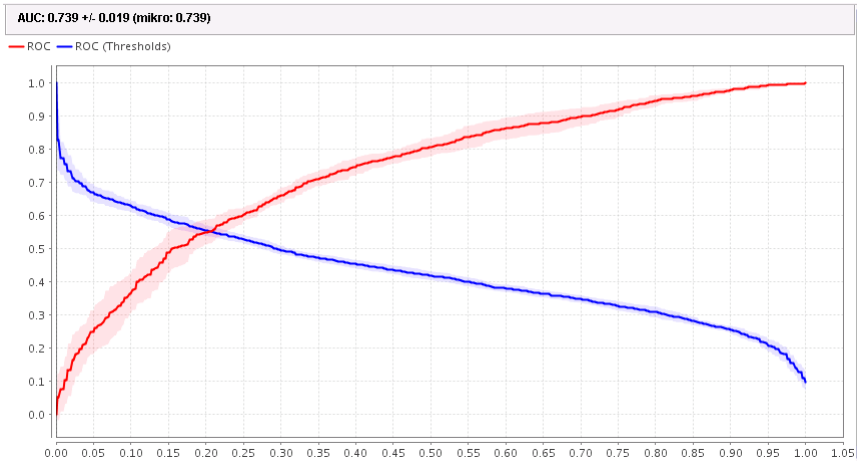

Figura C.12: ROC de Random Forest 1000 árvores

C.3 SVM RBF

C.3.1 Default

accuracy: 61.04% +/- 1.55% (mikro: 61.04%)

	true S	true N	class precision
pred. S	824	393	67.71%
pred. N	1146	1587	58.07%
class recall	41.83%	80.15%	

Figura C.13: Matriz de confusão de SVM RBF default

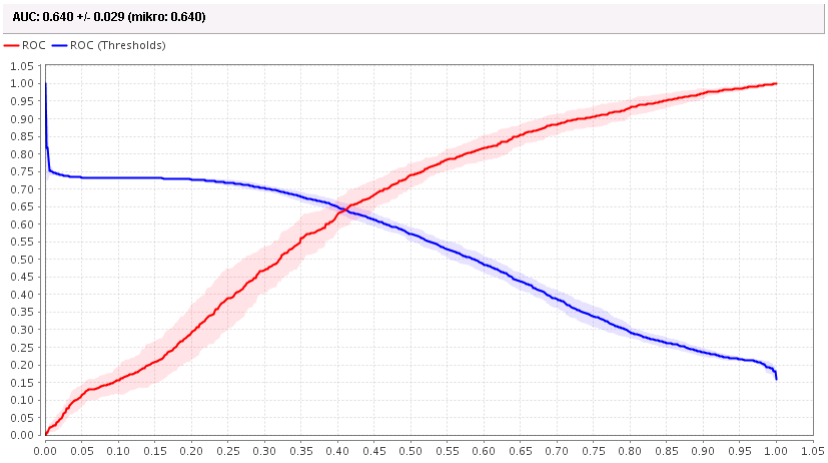

Figura C.14: ROC de SVM RBF default

C.3.2 (C=50 ;gamma=0.005)

accuracy: 62.96% +/- 1.44% (mikro: 62.96%)

	true S	true N	class precision
pred. S	1293	786	62.19%
pred. N	677	1194	63.82%
class recall	65.63%	60.30%	

Figura C.15: Matriz de confusão de SVM RBF (C=50 ;γ=0.005)

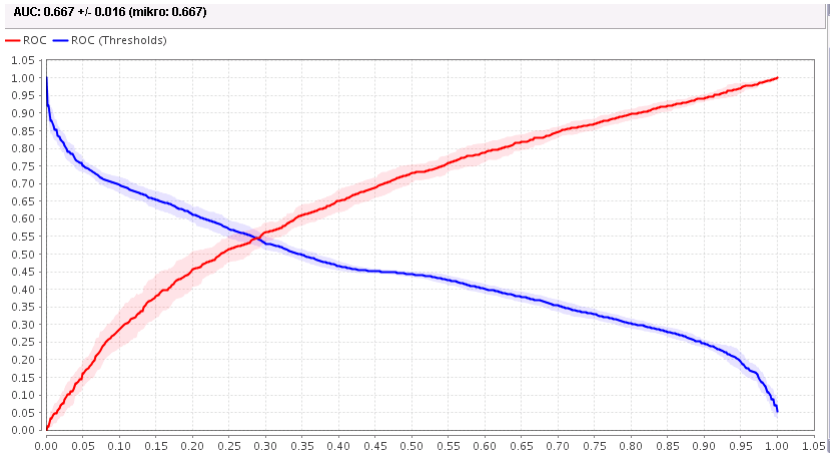

Figura C.16: ROC de SVM RBF (C=50 ;γ=0.005)

C.3.3 (1 ;gamma=0.75)

accuracy: 57.85% +/- 1.70% (mikro: 57.85%)

	true S	true N	class precision
pred. S	1827	1522	54.55%
pred. N	143	458	76.21%
class recall	92.74%	23.13%	

Figura C.17: Matriz de confusão de SVM RBF (1 ;γ=0.75)

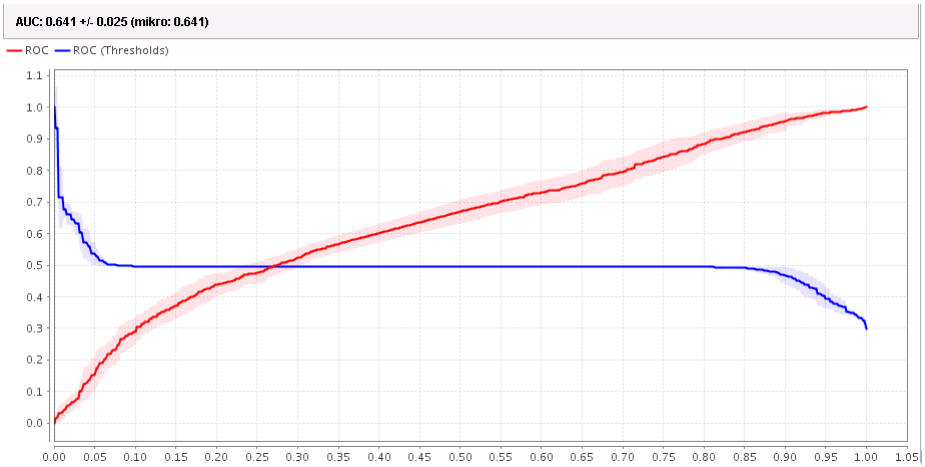

Figura C.18: ROC de SVM RBF (1 ;γ=0.75)

C.4 SVM Sigmoide

C.4.1 Default

accuracy: 53.97% +/- 3.11% (mikro: 53.97%)			
	true S	true N	class precision
pred. S	1085	933	53.77%
pred. N	885	1047	54.19%
class recall	55.08%	52.88%	

Figura C.19: Matriz de confusão de SVM sigmoide default

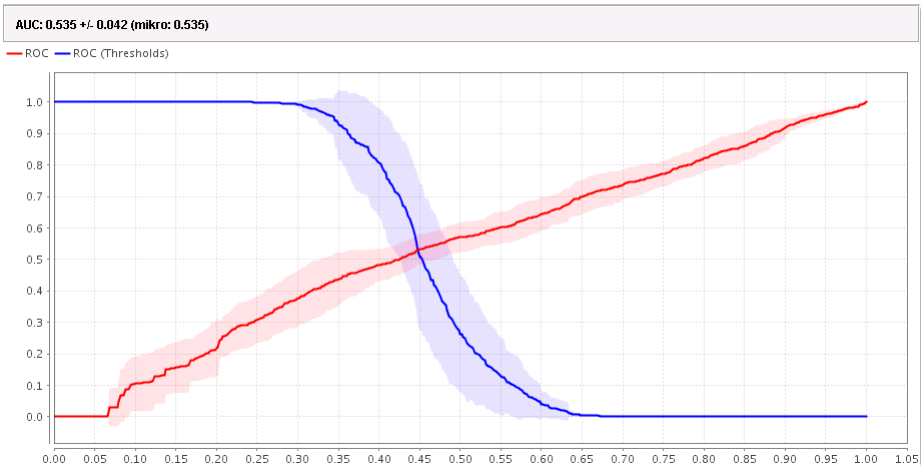

Figura C.20: ROC de SVM sigmoide default

C.4.2 C=200 ;gamma=10

accuracy: 54.30% +/- 5.60% (mikro: 54.30%)			
	true S	true N	class precision
pred. S	860	695	55.31%
pred. N	1110	1285	53.65%
class recall	43.65%	64.90%	

Figura C.21: Matriz de confusão de SVM sigmoide (C=200 ;γ=10)

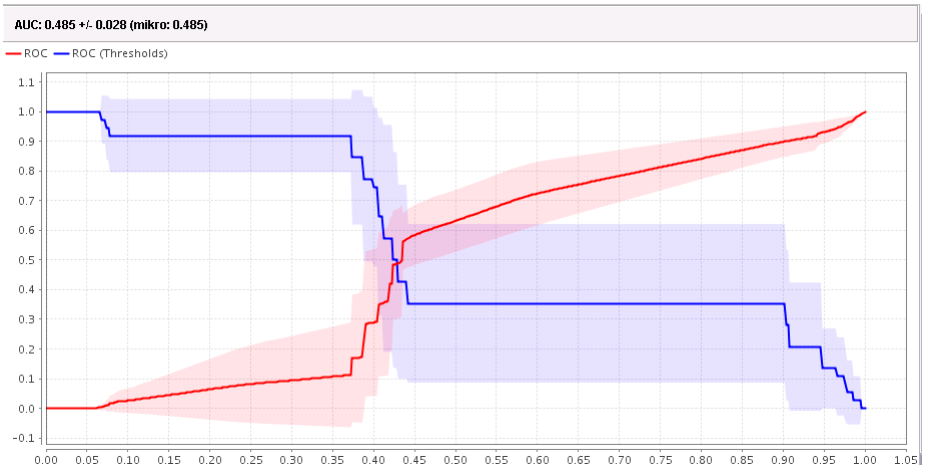

Figura C.22: ROC de SVM sigmoide (C=200 ;γ=10)

C.4.3 C=1 ;gamma=0.75

accuracy: 57.85% +/- 1.70% (mikro: 57.85%)

	true S	true N	class precision
pred. S	1827	1522	54.55%
pred. N	143	458	76.21%
class recall	92.74%	23.13%	

Figura C.23: Matriz de confusão de SVM sigmoide (C=1 ;γ=0.75)

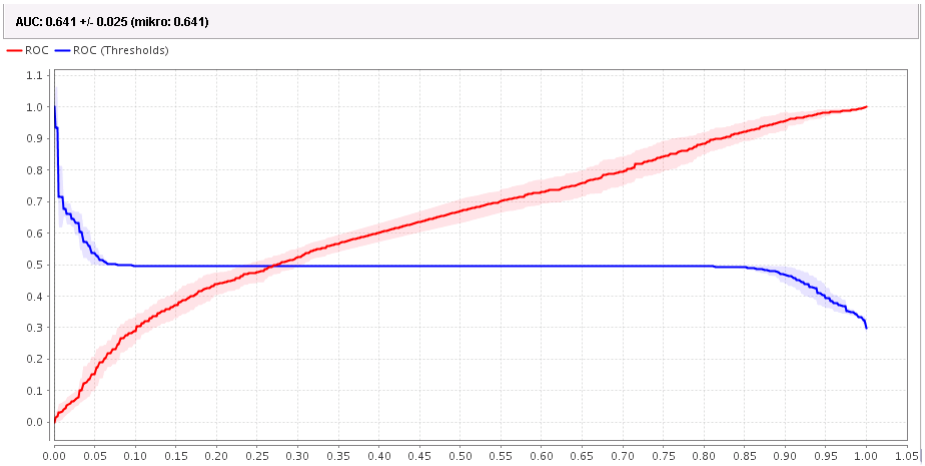

Figura C.24: ROC de SVM sigmoide (C=1 ;γ=0.75)

C.5 Teoria e resultados ILP

```
1 [theory]
```

```
2
3  [Rule 1] [Pos cover = 64 Neg cover = 10]
4  internamento(A) :-
5     consumo(A,B,C,D,E,F,'misturas de macronutrientes e micronutrientes',G,H,I,J,K,L)
          , lteq(G,200), estatistica(M,N,O,P,Q,R,S,T,U,V,W,A,X,Y,Z,A1,B1,C1
          ,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1).
6
7  [Rule 2] [Pos cover = 62 Neg cover = 7]
8  internamento(A) :-
9     consumo(A,B,C,D,E,F,'brometo de ipratropio+salbutamol',G,H,I,J,K,L), episodio(M,
          N,O,P,A,Q,R,S,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,
          Q1), gteq(N,28.0).
10
11 [Rule 3] [Pos cover = 27 Neg cover = 6]
12 internamento(A) :-
13    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1),
           lteq(C,19.0), gteq(C,19.0).
14
15 [Rule 4] [Pos cover = 49 Neg cover = 9]
16 internamento(A) :-
17    consumo(A,B,C,D,E,F,digoxina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,Y,
          Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), gteq(N,12.0).
18
19 [Rule 5] [Pos cover = 61 Neg cover = 8]
20 internamento(A) :-
21    consumo(A,B,C,D,E,F,'ac. acetilsalicilico',G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,
          S,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[77],M1,N1,O1,P1), gteq(
          N,17.0).
22
23 [Rule 6] [Pos cover = 35 Neg cover = 7]
24 internamento(A) :-
25    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,[77],B1,C1,D1,E1
          ), gteq(E1,90.0), estatistica(F1,G1,H1,1.0,I1,J1,K1,L1,M1,N1,O1,A,P1
          ,[98,114,111,109,101,116,111,32,100,101,32,105,112,114,97,116,114,243,112,
26    105,111,43,115,97,108,98,117,116,97,109,111,108],Q1,R1,S1,T1,U1,V1,W1).
27
28 [Rule 7] [Pos cover = 34 Neg cover = 10]
29 internamento(A) :-
30    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,[70],B1,C1,D1,E1
          ), lteq(C,0.0), estatistica(F1,G1,H1,0.0,I1,J1,K1,L1,M1,N1,O1,A,P1
          ,[100,101,115,99,111,110,104,101,99,105,100,111],Q1,R1,S1,T1,U1,V1,W1).
31
32 [Rule 8] [Pos cover = 37 Neg cover = 8]
33 internamento(A) :-
34    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,[77],B1,C1,D1,E1
          ), gteq(B,174.825012207031), estatistica(F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,A,
          Q1,R1,S1,T1,U1,V1,[100,101,115,99,111,110,104,101,99,105,100,111],W1,X1).
35
36 [Rule 9] [Pos cover = 43 Neg cover = 7]
```

```
37 internamento(A) :-
38    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1),
           gteq(C,16.0), estatistica(G1,H1,I1,250.0,J1,K1,L1,M1,N1,O1,P1,A,Q1
          ,[98,114,111,109,101,116,111,32,100,101,32,105,112,114,97,116,114,243,
39    112,105,111],R1,S1,T1,U1,V1,W1,X1).
40
41 [Rule 10] [Pos cover = 28 Neg cover = 8]
42 internamento(A) :-
43    consumo(A,B,C,D,E,F,paracetamol,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X
          ,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), gteq(P1,95.0).
44
45 [Rule 11] [Pos cover = 21 Neg cover = 7]
46 internamento(A) :-
47    estatistica(B,C,D,E,F,G,H,I,J,K,L,A,M,N,O,P,Q,R
          ,[100,101,115,99,111,110,104,101,99,105,100,111],S,T), lteq(H,4.5), gteq(H
          ,4.5).
48
49 [Rule 12] [Pos cover = 36 Neg cover = 6]
50 internamento(A) :-
51    consumo(A,B,C,D,E,F,vancomicina,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W,
          A,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1),
          lteq(S,25.0).
52
53 [Rule 13] [Pos cover = 20 Neg cover = 5]
54 internamento(A) :-
55    consumo(A,B,C,D,E,F,metoclopramida,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V
          ,W,A,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1),
          lteq(U,0.0).
56
57 [Rule 14] [Pos cover = 45 Neg cover = 7]
58 internamento(A) :-
59    consumo(A,B,C,D,E,F,'cloreto de potassio',G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S
          ,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,Q1), gteq(Q1
          ,90.0).
60
61 [Rule 15] [Pos cover = 36 Neg cover = 10]
62 internamento(A) :-
63    consumo(A,B,C,D,E,F,'ac. valproico',G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,
          V,W,A,X,Y,Z,A1,B1,C1,D1,E1,F1), gteq(R,2.83333333333333).
64
65 [Rule 16] [Pos cover = 25 Neg cover = 5]
66 internamento(A) :-
67    consumo(A,B,C,D,E,F,metilprednisolona,G,H,I,J,K,L), estatistica(M,N,O,1.0,P,Q,R,
          S,T,U,V,A,W
          ,[98,114,111,109,101,116,111,32,100,101,32,105,112,114,97,116,114,243,112
68    ,105,111,43,115,97,108,98,117,116,97,109,111,108],X,Y,Z,A1,B1,C1,D1), gteq(Q
          ,4.0).
69
70 [Rule 17] [Pos cover = 24 Neg cover = 6]
```

```
internamento(A) :-
   consumo(A,B,C,D,E,F,'ac. aminocaproico',G,H,I,J,K,L), gteq(C,3).

[Rule 18] [Pos cover = 40 Neg cover = 9]
internamento(A) :-
   consumo(A,B,C,D,E,F,esomeprazol,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W,
       A,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1),
       gteq(Q,1.52631578947368).

[Rule 19] [Pos cover = 16 Neg cover = 6]
internamento(A) :-
   episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,[80,111,114,116,117,103,97,108],S,T
       ,U,V,W,X,Y,Z,A1,B1,C1,D1,E1), lteq(B,4.72499990463257).

[Rule 20] [Pos cover = 24 Neg cover = 6]
internamento(A) :-
   cirurgia(A,B,C,D,E,F,G,H,I), consumo(A,J,K,L,M,N,'cloreto de sodio',O,P,Q,R,S,T)
       , lteq(O,500).

[Rule 21] [Pos cover = 35 Neg cover = 9]
internamento(A) :-
   consumo(A,B,C,D,E,F,carvedilol,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,
       Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), gteq(P1,85.0).

[Rule 22] [Pos cover = 29 Neg cover = 6]
internamento(A) :-
   consumo(A,B,C,D,E,F,cefazolina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,
       Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), gteq(N,21.0).

[Rule 23] [Pos cover = 25 Neg cover = 6]
internamento(A) :-
   consumo(A,B,C,D,E,F,dexametasona,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W
       ,A,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1),
       gteq(Q,1.2).

[Rule 24] [Pos cover = 29 Neg cover = 4]
internamento(A) :-
   consumo(A,B,C,D,E,F,sertralina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,
       Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,Q1), gteq(Q1,85.0).

[Rule 25] [Pos cover = 50 Neg cover = 5]
internamento(A) :-
   consumo(A,B,C,D,E,F,'piperacilina+tazobactam',G,H,I,J,K,L), episodio(M,N,O,P,A,Q
       ,R,S,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[77],M1,N1,O1,P1),
       gteq(P1,85.0).

[Rule 26] [Pos cover = 33 Neg cover = 10]
internamento(A) :-
```

```
108    consumo(A,B,C,D,E,F,ranitidina,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W,A
           ,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1), lteq
           (T,2.88461538461538).
109
110 [Rule 28] [Pos cover = 39 Neg cover = 10]
111 internamento(A) :-
112    consumo(A,B,C,D,E,F,lorazepam,G,H,I,J,K,L), estatistica(M,N,O,1.0,P,Q,R,S,T,U,V,
           A,W,[98,114,111,109,101,116,111,32,100,101,32,105,112,
113 114,97,116,114,243,112,105,111,43,115,97,108,98,117,116,97,109,111,108],X,Y,Z,A1,B1
       ,C1,D1).
114
115 [Rule 29] [Pos cover = 30 Neg cover = 10]
116 internamento(A) :-
117    consumo(A,B,C,D,E,F,losartan,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W,A,X
           ,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1), gteq(R
           ,3.0).
118
119 [Rule 30] [Pos cover = 22 Neg cover = 9]
120 internamento(A) :-
121    consumo(A,B,C,D,E,F,cloropromazina,G,H,I,J,K,L), gteq(B,1), episodio(M,N,O,P,A,Q
           ,R,S,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[77],M1,N1,O1,P1).
122
123 [Rule 31] [Pos cover = 26 Neg cover = 7]
124 internamento(A) :-
125    consumo(A,B,C,D,E,F,tramadol,G,H,I,J,K,L), estatistica(M,N,O,10.0,P,Q,R,S,T,U,V,
           A,W,[109,101,116,111,99,108,111,112,114,97,109,105,100,97],X,Y,Z,A1,B1,C1,D1
           ), lteq(U,0.0).
126
127 [Rule 32] [Pos cover = 10 Neg cover = 2]
128 internamento(A) :-
129    consumo(A,B,C,D,E,F,petidina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,Y,
           Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), lteq(N,1.0).
130
131 [Rule 33] [Pos cover = 13 Neg cover = 6]
132 internamento(A) :-
133    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,[80,111,114,116,117,103,97,108],S,T
           ,U,V,W,X,Y,Z,A1,B1,C1,D1,E1), gteq(E1,90.0), estatistica(F1,G1,H1,0.0,I1,J1,
           K1,L1,M1,N1,O1,A,P1,[100,101,115,99,111,110,104,101,99,105,100,111],Q1,R1,S1
           ,T1,U1,V1,W1).
134
135 [Rule 34] [Pos cover = 18 Neg cover = 0]
136 internamento(A) :-
137    cirurgia(A,B,C,D,E,[100,101,115,99,111,110,104,101,99,105,100,111],F,G,H),
           estatistica(I,J,K,L,M,N,O,P,Q,R,S,A,T,U,V,W,X,Y,Z,A1,B1), lteq(R,0.0).
138
139 [Rule 35] [Pos cover = 62 Neg cover = 10]
140 internamento(A) :-
```

```
141    consumo(A,B,C,D,E,F,'piperacilina+tazobactam',G,H,I,J,K,L), estatistica(M,N,O,P,
           Q,R,S,T,U,V,W,A,X,Y,Z,A1,B1,C1
           ,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1), gteq(S,250.0).
142
143 [Rule 36] [Pos cover = 18 Neg cover = 5]
144 internamento(A) :-
145    consumo(A,B,C,D,E,F,nifedipina,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W,A
           ,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1).
146
147 [Rule 37] [Pos cover = 15 Neg cover = 10]
148 internamento(A) :-
149    consumo(A,B,C,D,E,F,'electrolitos+glucose',G,H,I,J,K,L), estatistica(M,N,O,P,Q,R
           ,S,T,U,V,W,A,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],
           D1,E1), lteq(T,1.0).
150
151 [Rule 38] [Pos cover = 25 Neg cover = 9]
152 internamento(A) :-
153    estatistica(B,C,D,E,F,G,H,I,J,K,L,A,M,N,O,P,Q,R,S,T,U), lteq(F,0.9), gteq(J,0.9)
           .
154
155 [Rule 39] [Pos cover = 25 Neg cover = 10]
156 internamento(A) :-
157    consumo(A,B,C,D,E,F,diclofenac,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,
           Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[77],M1,N1,O1,P1), gteq(N,14.0).
158
159 [Rule 40] [Pos cover = 34 Neg cover = 5]
160 internamento(A) :-
161    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,[77],B1,C1,D1,E1
           ), lteq(E1,50.0), gteq(B,165.375).
162
163 [Rule 41] [Pos cover = 27 Neg cover = 7]
164 internamento(A) :-
165    consumo(A,B,C,D,E,F,levofloxacina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W
           ,X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[77],M1,N1,O1,P1), gteq(P1,60.0).
166
167 [Rule 42] [Pos cover = 56 Neg cover = 10]
168 internamento(A) :-
169    consumo(A,B,C,D,E,F,meropenem,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,Y
           ,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[77],M1,N1,O1,P1), estatistica(Q1,R1,
           S1,T1,U1,V1,W1,X1,Y1,Z1,A2,A,B2,C2,D2,E2,F2,G2
           ,[100,101,115,99,111,110,104,101,99,105,100,111],H2,I2).
170
171 [Rule 43] [Pos cover = 54 Neg cover = 8]
172 internamento(A) :-
173    consumo(A,B,C,D,E,F,metformina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,
           Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,Q1), gteq(N,8.0).
174
175 [Rule 44] [Pos cover = 18 Neg cover = 5]
176 internamento(A) :-
```

```
177    estatistica(B,C,D,1000.0,E,F,G,H,I,J,K,A,L
           ,[112,97,114,97,99,101,116,97,109,111,108],M,N,O,P,Q,R,S), lteq(F,5.0), gteq
           (H,6.0).
178
179 [Rule 45] [Pos cover = 41 Neg cover = 9]
180 internamento(A) :-
181    consumo(A,B,C,D,E,F,bisoprolol,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W,A
           ,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1), lteq
           (S,2.5).
182
183 [Rule 46] [Pos cover = 45 Neg cover = 6]
184 internamento(A) :-
185    consumo(A,B,C,D,E,F,domperidona,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,W,
           A,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1),
           gteq(U,0.0075).
186
187 [Rule 47] [Pos cover = 19 Neg cover = 5]
188 internamento(A) :-
189    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,[70],B1,C1,D1,E1
           ), gteq(E1,85.0), estatistica(F1,G1,H1,10.0,I1,J1,K1,L1,M1,N1,O1,A,P1
           ,[109,101,116,111,99,108,111,112,114,97,109,105,100,97],Q1,R1,S1,T1,U1,V1,W1
           ).
190
191 [Rule 48] [Pos cover = 15 Neg cover = 6]
192 internamento(A) :-
193    estatistica(B,C,D,E,F,G,H,I,J,K,L,A,M,N,O,P,Q,R
           ,[100,101,115,99,111,110,104,101,99,105,100,111],S,T), lteq(G,1.0), gteq(I
           ,9999.0).
194
195 [Rule 50] [Pos cover = 25 Neg cover = 6]
196 internamento(A) :-
197    consumo(A,B,C,D,E,F,G,H,I,J,K,L,M), gteq(H,75), estatistica(N,O,P,20.0,Q,R,S,T,U
           ,V,W,A,X,[102,117,114,111,115,101,109,105,100,97],Y,Z,A1,B1,C1,D1,E1).
198
199 [Rule 52] [Pos cover = 15 Neg cover = 4]
200 internamento(A) :-
201    estatistica(B,C,D,5.0,E,F,G,H,I,J,K,A,L
           ,[100,101,120,97,109,101,116,97,115,111,110,97],M,N,O,P,Q,R,S), lteq(J,0.0).
202
203 [Rule 53] [Pos cover = 29 Neg cover = 9]
204 internamento(A) :-
205    consumo(A,B,C,D,E,F,alprazolam,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,
           Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), gteq(N,9.0).
206
207 [Rule 55] [Pos cover = 23 Neg cover = 9]
208 internamento(A) :-
209    cirurgia(A,B,C,D,E,[108],F,G,H), episodio(I,J,K,L,A,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,
           A1,B1,C1,D1,E1,F1,G1,H1,[77],I1,J1,K1,L1), gteq(L1,85.0).
210
```

```
211 [Rule 58] [Pos cover = 69 Neg cover = 10]
212 internamento(A) :-
213    consumo(A,B,C,D,E,F,sinvastatina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,
           X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), gteq(N,14.0).
214
215 [Rule 62] [Pos cover = 18 Neg cover = 8]
216 internamento(A) :-
217    estatistica(B,C,D,15.0,E,F,G,H,I,J,K,A,L,[108,97,99,116,117,108,111,115,101],M,N
           ,O,P,Q,R,S), gteq(E,1.0).
218
219 [Rule 65] [Pos cover = 14 Neg cover = 8]
220 internamento(A) :-
221    consumo(A,B,C,D,E,F,varfarina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,Y
           ,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[70],M1,N1,O1,P1), lteq(N,7.0).
222
223 [Rule 75] [Pos cover = 29 Neg cover = 10]
224 internamento(A) :-
225    consumo(A,B,C,D,E,F,fitomenadiona,G,H,I,J,K,L), estatistica(M,N,O,P,Q,R,S,T,U,V,
           W,A,X,Y,Z,A1,B1,C1,[100,101,115,99,111,110,104,101,99,105,100,111],D1,E1),
           gteq(S,10.0).
226
227 [Rule 104] [Pos cover = 27 Neg cover = 8]
228 internamento(A) :-
229    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z,A1,[70],B1,C1,D1,E1
           ), gteq(B,161.875), estatistica(F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,A,Q1,R1,S1,
           T1,U1,V1,[100,101,115,99,111,110,104,101,99,105,100,111],W1,X1).
230
231 [Rule 107] [Pos cover = 24 Neg cover = 10]
232 internamento(A) :-
233    consumo(A,B,C,D,E,F,tramadol,G,H,I,J,K,L), estatistica(M,N,O,1000.0,P,Q,R,S,T,U,
           V,A,W,[112,97,114,97,99,101,116,97,109,111,108],X,Y,Z,A1,B1,C1,D1), lteq(U
           ,3000.0).
234
235 [Rule 112] [Pos cover = 19 Neg cover = 10]
236 internamento(A) :-
237    consumo(A,B,C,D,E,F,'gluconato de calcio',G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S
           ,T,U,V,W,X,Y,Z,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,M1,N1,O1,P1,Q1), lteq(Q1
           ,70.0).
238
239 [Rule 125] [Pos cover = 14 Neg cover = 8]
240 internamento(A) :-
241    episodio(B,C,D,E,A,F,G,H,I,J,K,L,M,N,O,P,Q,R,[80,111,114,116,117,103,97,108],S,T
           ,U,V,W,X,Y,Z,A1,B1,C1,D1,E1), lteq(E1,15.0), gteq(C,15.0).
242
243 [Rule 189] [Pos cover = 31 Neg cover = 8]
244 internamento(A) :-
245    consumo(A,B,C,D,E,F,tiamina,G,H,I,J,K,L), episodio(M,N,O,P,A,Q,R,S,T,U,V,W,X,Y,Z
           ,A1,B1,C1,D1,E1,F1,G1,H1,I1,J1,K1,L1,[77],M1,N1,O1,P1), gteq(P1,70.0).
246
```

```
[Training set performance]
            Actual
        +          -
     + 1101        342        1443
Pred
     - 672        1440        2112

       1773        1782        3555

Accuracy = 0.714767932489451
[Training set summary] [[1101,342,672,1440]]
[Test set performance]
           Actual
        +          -
     + 40        19        59
Pred
     - 157        179        336

       197        198        395

Accuracy = 0.554430379746835
[Test set summary] [[40,19,157,179]]
[time taken] [314346.778]

[total clauses constructed] [4886493]
```